THE UNIVERSITY
of EDINBURGH

The Evolutionary Origin, Cyclic Peptide Targeting and Physiological Activation of Peptidyl Arginine Deiminase 4

Thomas Cummings

Thesis Presented for the Degree of Doctor of Philosophy
The University of Edinburgh
October 2019

Declaration

I hereby declare that the work presented in this PhD thesis is the original work of the author, except where specific reference is made to other sources. This thesis has been composed solely by myself and has not been submitted for any other degree, diploma or qualification.

Thomas Cummings
14th October 2019
Corrections 1st April 2020

Acknowledgements

Firstly, I would like to thank all the members of the Christophorou lab for their help and support including David Hay, Abby Wilson, Gavriil Gavriilidis, Emma Clarke and Christine Young. I would also like to thank Dr Maria Christophorou.

I was lucky to have other people around who provided inspiration and motivation. I am particularly indebted to Chris Ponting for getting me over the line and for all of his advice and wisdom along the way. I am also grateful for all of his intellectual contributions to the Evolution chapter and for pushing me on to produce this thesis. I'd also like to thank Louise Walport for such an exciting collaboration with the peptides and for all her support. I look forward to seeing all the amazing work that will come out of her lab.

I'd like to thank the various scientific collaborators without whom much of this project would not have been possible. In particular, Hiroaki Suga from the University of Tokyo for overseeing and supervising the peptide work; Greg Findlay, Philip Cohen, and Dario Alessi from the MRC PPU, and Nathanael Gray from Harvard University for their help with phosphorylation and signalling networks and for providing a kinase inhibitor library; Ana Mendanha Falcao, Mandy Meijer and Goncalo Castelo Branco from the Karolinska Institutet for a collaboration with PADI2beta that did not quite make it into the thesis; and Clive D'Santos from the University of Cambridge for his help with mass spectrometry.

Closer to home, I would like to thank David Dorward and Adriano Rossi for their help with neutrophil biology at the QMRI; Andy Finch, Noor Gammoh, Alex von Kriegsheim, Juan Carlos Acosta, Tamir Chandra and all present and past members of our big group lab meetings for their support and scrutiny, and to Priya Hari for helping me with high content microscopy. I'd like to thank all the members of the ES cell culture room, members of Genome Regulation section, Lizzie Fryer for support with FACS, and

Technical Services for all their help preparing reagents. Special thanks go to Jimi Wills and Niall Quinn for their help with mass spectrometry and to instructors at the Biochemical Society course on Quantitative mass spectrometry. Thanks also go to the brilliant instructors from the MadPhylo course in Madrid including John Huelsenbeck, Sebastian Höhna, Brian Moore and Michael May for all their help with phylogenetic analysis; to Luis Sanchez Pulido for his insight and advice with computational and protein analysis; and to John Ireland for showing me how to speak to Eddie in UNIX.

I'm hugely grateful to my friends and family. Thank you to all the other people in my PhD cohort including Issy MacGregor and Dan Dodd, with particular thanks to Yan Ping Lee and Victòria Gudiño. Amongst other people, I would like to thank Hugh Oldham, John Gregg, Alexander Bell, William Turner, Ivan Voynitsky, Ashley Spinelli, Alison Bissell, Greta Alijošiūtė, Hannah Borhan-Azad, Zoltan Gera, Martyna Gražiūnaitė, Alice Fiennes, Indigo Bates, Daisy Harvey, Robert Adam, Barriss Offee, Robert Natzler, Henry and Ashley Dee, and Merle Robbins.

In particular, I'd like to thank my grandparents for providing so much inspiration; my mother and father for all their love and support; and my brother Daniel for keeping me sane. Last but not least, thanks to Molly, who knows more about any of this than she ever would have wanted, and without whom this thesis would not have been finished.

'I am in blood / Stepp'd in so far that, should I wade no more, / Returning were as tedious as go o'er'

Abstract

Five peptidylarginine deiminases (PADIs), regulating diverse aspects of human physiology, catalyse the post-translational modification citrullination in mammals. Deregulation of citrullination underlies the development of diverse pathologies and PADI4 is of particular medical interest in autoimmune diseases and in cancer. PADI activation is understood biochemically at the level of the crystal structure where allosteric calcium ion binding drives an extensive conformational change and the nucleophilic cysteine relocates approximately 5-10 Å. The critical open question for the field is how this activation can occur in cells. The calcium concentrations required to drive the conformational change in vitro are supraphysiological by two orders of magnitude and little is understood mechanistically about how cellular signalling events enable PADI enzyme activation. As such, the biological activation and therefore the deregulation of PADIs in diseases have remained enigmatic. This PhD set out to understand the physiological activation and regulation of the PADI4 enzyme.

First, I identified a surprising origin of the mammalian PADI enzyme family as deriving from an ancient horizontal gene transfer event from cyanobacteria. A variety of maximum likelihood and Bayesian phylogenetic methods as well as non-parametric protein sequence analyses revealed that animal PADIs did not derive from the last universal common ancestor by conventional vertical descent and instead represent an ancient horizontal acquisition. The cyanobacterial PADI is shown to be active with a more stringent calcium dependency than the mammalian counterpart. This has particular interest as to how mechanisms for physiological activation of the human enzyme may have evolved and provides a rare example of HGT into the human lineage.

Then, tractable conditions were established for comparing resting PADI4 to activated PADI4 across a variety of cell systems relevant to innate immunity and pluripotency. Emphasis was placed on identifying cellular stimuli that may be independent of calcium flux or transcriptional changes and in which

the levels of PADI4 protein may be stringently controlled and independent from the activating stimulus. In this model system, PADI4 was shown to be activated by canonical Wnt signalling – downstream of recombinant Wnt3a, short-term GSK3β inhibition and chemical Wnt agonism. PADI4 inhibition reduced Wnt-driven transcription in a reporter assay. A role for PADI4 in Wnt signalling would be important for roles in haematopoetic stem cells, in neutrophils, in the establishment of pluripotency and its overexpression in cancer. Future work will be required to confirm these exciting effects in a physiologically relevant context.

Next, work was undertaken to screen a vast library of peptide molecules to find candidates that bind with high affinity to different PADI4 conformations, using the RaPID system, performed in collaboration with Dr Walport and Prof Suga. First, the active calcium-soaked conformation of PADI4 was targeted to find active site inhibitors. Second, the active site was covalently blocked to identify possible activating peptides. Biochemical and cell biological assays revealed different peptides that bound tightly to both conformations of PADI4 (low nM binding affinity) and a biotinylated peptide was characterized for pulldown from lysates. Different peptides were shown both to inhibit and activate PADI4 with efficacy in vitro and in live cells. A toolkit of PADI4 specific cyclic peptide reagents can therefore be taken forward to address new biological questions. The activator in particular represents the first such cyclic peptide epigenetic activator.

Finally, conditions for the cellular activation of PADI4 were used in high-resolution proteomic analysis of PADI4 after affinity pulldown. Conditions were optimized to pulldown PADI4 using the new biotinylated cyclic peptide reagent developed in Chapter 5 and a commercial antibody. An irreversible inhibitor was used to distinguish allosteric interactors from substrates. Mass spectrometry showed that a number of proteins with established calcium-binding functions (EF-hand containing proteins), including calmodulin, were enriched for allosteric binding to the inactive conformation of PADI4. The

interaction, which was enriched in the absence of calcium, was confirmed by reciprocal pulldown of calmodulin. In vitro experiments showed that recombinant calmodulin activates PADI4 at a reduced calcium concentration. Putative calmodulin binding sites were identified on PADI4 using bioinformatic analysis. These can be readily mutated by site-directed mutagenesis and PADI4 variants introduced into cells for future validation of the effect in cells. This work represents clear progress towards understanding a complete mechanism for PADI4 activation in cells.

Lay Abstract

Most people are familiar with the idea of DNA as the genetic code for creating life. What is less well known, is the way in which this genetic information is converted into instructions for the cell. Triplets of DNA bases, called codons, code for just 20 amino acids, which are connected together in long strings to make proteins. After proteins have been made, however, they can be further modified in the cell in a controlled way. These are called post-translational modifications and are a fundamental way for cells to send and receive signals and control activity.

This study examines the role of the PADI enzymes, and in particular PADI4. PADI4 can be found in white blood cells and in stem cells. This enzyme is responsible for turning arginine (one of the coded amino acids) into citrulline (one of the modifications made to existing proteins in real time). What has been observed is that excessive citrullination is present in various disease contexts. In particular, the over-expression of PADI proteins (and therefore the over-production of citrullinated proteins) is implicated in many diseases including Rheumatoid Arthritis, Multiple Sclerosis, and some cancers. Several experiments have shown that PADIs can be activated using calcium in vitro, but this process requires far more calcium than is present in live cells. How the process is switched on and why it goes wrong in disease contexts is therefore poorly understood.

This research explores the regulation of PADI4 from several angles and using diverse methodologies:

Chapter 3 demonstrates that the gene, which codes for the PADI enzyme, originated in ancient bacteria and not in vertebrates as was previously thought. Instead of being inherited from parent to offspring over billions of years, the gene instead jumped from bacteria into early marine ancestors of animals (this explains its absence from the genome of many eukaryotes such as plants, yeast, or amoeba). By cloning the ancient gene and synthesizing

the protein in bacteria, I show that the cyanobacterial PADI is active, but with a more stringent calcium dependency than the mammalian enzyme. This has particularly interesting implications for how mechanisms for the activation of the human enzyme may have evolved.

In Chapter 4, I set up conditions in a variety of cell systems so that PADI4 can be activated– including in white blood cells, stem cells, and a cancer cell line. I showed that PADI4 was activated by a broader, biologically significant cell signaling pathway (called Wnt signalling) in one of these cell systems.

Chapter 5 describes a collaboration I initiated with a former colleague to target PADI4 to be able to modify its activity with cyclic peptides. It uses novel screening techniques to identify new synthetic cyclic peptide molecules that can inhibit PADI4, extract it from cells, or activate it. These molecules will provide a toolkit of reagents that can help enable the precise role of PADI4 in biological settings to be studied.

In Chapter 6, I use high-resolution mass spectrometry to try to find other proteins that bind with PADI4 in cells and which may differ between resting and activated PADI4. I generated a dataset of regulatory candidates that provide a platform for elucidating the physiological activation of PADI4 in cells. I then tested one of the most promising of these candidates showing that a calcium-binding protein called calmodulin interacts with PADI4 and activates the enzyme at lower calcium *in vitro*. This can be explored in more mechanistic detail in future work.

Table of Contents

List of Figures

Chapter 1: Introduction

Chapter 2: Materials and Methods

Chapter 3: The evolutionary origin of the peptidyl arginine deiminases

Chapter 4: Activating PADI4 in Cells

Chapter 5: Cyclic peptide reagents to target PADI4 for inhibition, activation and affinity purification

Chapter 6: Towards the Physiological Activation of PADI4

List of Abbreviations

Standard three letter amino acid abbreviations are used throughout.

A

ABAP: Antibody Based Assay for PADI Activity

ACPAs: Anti-citrullinated Protein Antibodies

ADI: Free Arginine Deiminase

AGAT: Glycine Amidinotransferase

AgD: Agmatine Deiminase

AGD: Accumulated Genetic Divergence

ACN: Acetonitrile

aLRT: Approximate Likelihood Ratio Test

AMBIC: Ammonium bicarbonate

APS: Ammonium Persulfate

ASL: Argininosuccinate Lyase

ASS: Argininosuccinate Synthase

ATRA: All-trans-retinoic Acid

ATP: Adenosine Triphosphate

AU: Approximately Unbiased

B

BAEE: Nα-benzoyl-l-arginine ethyl ester

BSA: Bovine Serum Albumin

BS3: Bis(sulfosuccinimidyl) suberate

C

CaM: Calmodulin

CETSA: Cellular Thermal Shift Assay

ChIP: Chromatin Immunoprecipitation

CK1alpha: Casein Kinase 1 alpha

COLDER: Colour Developing Reagent

CRAP-ome: Contaminant Repository for Affinity Purification Mass Spectometry

CV: Column Volumes

D

DDAH: Dimethylarginine Dimethylaminohydrolase

DEA NONOate: Diethylammonium (Z)-1-(N,N-diethylamino)diazen-1-ium-1,2-diolate

DNA: Deoxyribonucleic acid

DAPI: 4′,6-diamidino-2-phenylindole

DMSO: Dimethyl Sulfoxide

DTT: Dithiothreitol

E

E. coli: Escherichia coli

EAE: Experimental Autoimmune Encephalomyelitis

EDTA: Ethylenediaminetetraacetic acid

EGT: Endosymbiotic Gene Transfer

ELISA: Enzyme- linked Immunosorbent Assay

ELW: Expected Likelihood Weight

ERAD: Endoplasmic Reticulum Associated Protein Degradation

ES: Embryonic Stem

ESS: Effective Sample Size

F

FACS: Fluorescence-Activated Cell Sorting

FAM: Fluorescein amidite

FCS: Fetal Calf Serum

FIT: Flexizyme in vitro Translation

fMLF: N-Formyl-Met-Leu-Phe chemotactic peptide

FPLC: Fast protein liquid chromatography

FUS: Fused in Sarcoma

G

gADI: Arginine Deminase from *Giardia lamblia*

GFP: Green Fluorescent Protein

GLDH: Glutamate Dehydrogenase

GMEM: Glasgow's Modified Eagle's Minimum Essential Medium

GSK3: Glycogen Synthase Kinase 3

GST: Glutathione S-transferase

GST-His: Glutathione S-transferase – 6xHistidine

GWAS: Genome Wide Association Studies

H

HEPES: 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid

HGT: Horizontal Gene Transfer

HIF: Hypoxia Inducible Factor

HGU: Human Genetics Unit

HMM: Hidden Markov Model

HSC: Haematopoietic Stem Cell

I

IC50: Half Maximal Inhibitory Concentration

Ig: Immunoglobulin

IGMM: Institute of Genetics and Molecular Medicine

IMDM: Iscove's Modified Dulbecco's Medium

ING4: Inhibitor Of Growth Family Member 4

IP: Immunoprecipitation

iPTG: Isopropyl β-D-1-thiogalactopyranoside
iPS: Induced Pluripotent Stem
ITRs: Inverted Terminal Repeat Sequences

K

KSR: KnockOut Serum Replacement

L

LIF: Leukaemia inhibitory factor
LB: Luria-Bertani
LFQ: Label-free quantification
LPS: Lipopolysaccharides
LTA: Lipoteichoic Acid
LUCA: Last Universal Common Ancestor

M

mAbs: Mouse Antibodies
MBP: Myelin Basic Protein
MCMC: Markov chain Monte Carlo
MEK: Mitogen-activated protein kinase
MHC: Major Histocompatibility Complex
MMTS: Methyl-methanethiosulfonate
Mod-Cit: Antibody to Chemically Modified Citrulline
MPO: Myeloperoxidase
mRNA: Messenger RNA
MS: Multiple Sclerosis
MS/MS: Tandem Mass Spectrometry
m/z: Mass-to-charge ratio

N

NAWH: normal appearing white matter
NCBI: National Centre for Biotechnology Information
NET: Neutrophil Extra-cellular Trap
NF-κB p65: p65 subunit of nuclear factor κB
NLS: Nuclear Localization Signal
NO: Nitrous Oxide

O

OD600: Optical density at a wavelength of 600 nm
OTC: Ornithine Transcarbamylase

P

PADI: Peptidylarginine Deiminase
PAGE: Polyacrylamide Gel Electrophoresis
PBS: Phosphate-buffered saline
PCR: Polymerase chain reaction
PKC: Protein Kinase C
PMA: Phorbol 12-Myristate 13-Acetate

PMSF: Phenylmethylsulfonyl Fluoride

pPAD: porphyromonas-type peptidylarginine deiminase from *Porphyromonas gingivalis*

PRP: Platelet-Rich Plasma

PPU: MRC Protein and Phosphorylation Unit

PSI-BLAST: Position-Specific Iterated BLAST

PSRF: Potential Scale Reduction Factor

PTM: Post-Translational Modifications

Q

QMRI: Queen's Medical Research Institute

R

RA: Rheumatoid Arthritis

RaPID: Random nonstandard Peptides Integrated Discovery

RIME: Rapid immunoprecipitation mass spectrometry of endogenous proteins

RPMI: Roswell Park Memorial Institute

RiPP: Ribosomally Synthesized and Post-translationally Modified Peptide

RNA: Ribonucleic acid

RNAP2: RNA polymerase II

S

SAR: Structure-Activity Relationship

SDS: Sodium Dodecyl Sulfate

SH: Shimodaira-Hasegawa

STAGE: Stop-and-go-extraction

SNP: Single Nucleotide Polymorphism

SPPS: Solid Phase Peptide Synthesis

SPR: Surface Plasmon Resonance

T

TBS: Tris-buffered Saline

TBS-T: Tris-buffered Saline with 0.5% Tween-20 detergent

TBS-lo-T: Tris-buffered Saline with 0.1% Tween-20 detergent

TFA: Trifluoroacetic Acid

TFP: Trifluoperazine

TNFalpha: Tumour Necrosis Factor alpha

TOP-GFP: Cignal TCF/LEF Reporter assay

tRNA: Transfer ribonucleic acid

U

UCED: Uncorrelated exponentially distributed

UCLN: Uncorrelated lognorma

Chapter 1: Introduction

1.1 Epigenetic landscape

C.H. Waddington's picture of an epigenetic landscape is one of the most famous visual metaphors in biology[1]. A totipotent cell, drawn as a ball at the top of a hill, rolls down an increasingly restricted choice of grooves to a point of terminal differentiation, when it comes to rest at the end of the valley[1]. Waddington's elegant definition of the field of epigenetics – the study of "the causal interactions between genes and their products which bring the phenotype into being" – has made way for a modern (and in some ways quite clumsy) definition that stresses heritability outside of sequence changes in Deoxyribonucleic acid (DNA) and removes the emphasis on a mechanistic understanding that might link genes to their products[1-3].

Two remarkable experiments have transformed our conceptual understanding of cell specification by defying gravity in Waddington's metaphor. Taken together, they also reveal the importance of his original definition of epigenetics as a natural mode of progression in the field from genetics, as one that addresses the central question as to how a cell can exhibit such phenotypic diversity from the same genetic information. The first of these remarkable experiments was in 1958, when John Gurdon successfully cloned a frog[4]. The experiment, behind the cartoonish quality, answers a profound "epigenetic" question. In transferring the genetic content of a differentiated somatic cell (its nucleus) to an unfertilized egg and in so doing creating a viable frog, it showed that the differentiated cell's DNA contains all the information required to produce any cell in the cloned frog- no theoretical barriers exist or loss of information occurs during the process of differentiation[4]. In 2006, a second classic experiment (Takahashi and Yamanaka) stretched the logic of the first to surprising and almost absurd simplicity[5]. On provision of just four proteins, somatic cells were reprogrammed into pluripotent stem cells, with the capacity to redifferentiate back down any lineage[5]. In the language of Waddington's definition, this

showed that the subsequent causal interactions of merely four gene products on a differentiated cell's genes are sufficient to reprogram the cell completely. From this, it is perhaps not altogether surprising that somatic cells can be converted directly into another somatic cell state without returning to pluripotency: more recently four gene products (Brn2, Ascl1, Myt1l, Ngn2) convert T cells directly into functional neurons in just a few days[6].

Two further discoveries set the scene by placing the enzymes that form the study of this thesis, the tissue-restricted and stringently regulated peptidyl arginine deiminases (PADIs or PADs), at the two extremes of Waddington's landscape. On the one hand, PADI4 is highly expressed in terminally differentiated granulocytes (marbles resting in the valley); its activity is then switched on in neutrophil extracellular trap formation (NET formation or NETosis), a cell death program used as an innate immune defence mechanism by neutrophils in response to certain infectious agents[7,8]. On the other hand, PADI4 is expressed and its activity switched on in the Yamanaka experiment, during the process of reprogramming somatic cells to induced pluripotent stem cells (marbles defying gravity and traversing back to the top of Waddington's hill)[9].

1.2 From genetics to post translational modifications

Efforts to put a number to the human gene products contained in the genome have been debated extensively; the number has gradually reduced even since the publication of the draft sequence of the human genome[10]. This number is beginning to settle on something close to 42000, with just under half of those being protein-coding genes[11,12]. Of course, further diversity and complexity is achieved in other ways than possessing more and more genes (Figure 1.1). This includes post-transcriptional mechanisms such as splicing where segments of genes can be aggregated combinatorially to produce alternative isoforms before translation (Figure 1.1). It also extends to post-translational mechanisms, such as the huge variety of chemical modifications

of individual amino acids (post-translational modifications or PTMs) added to protein sequences after they have been translated, which thereby expand the genetic code (Figure 1.1). Many gene products code for proteins that catalyze the addition of PTMs to other proteins, but also catalyse PTMs to DNA, Ribonucleic acid (RNA) and gene regulatory proteins (such as histone tails or transcription factors) that can affect subsequent expression. PTMs provide a crucial "epigenetic" link (once again to use Waddington's sense of the term) to explain the interactions between genes and their products in both directions.

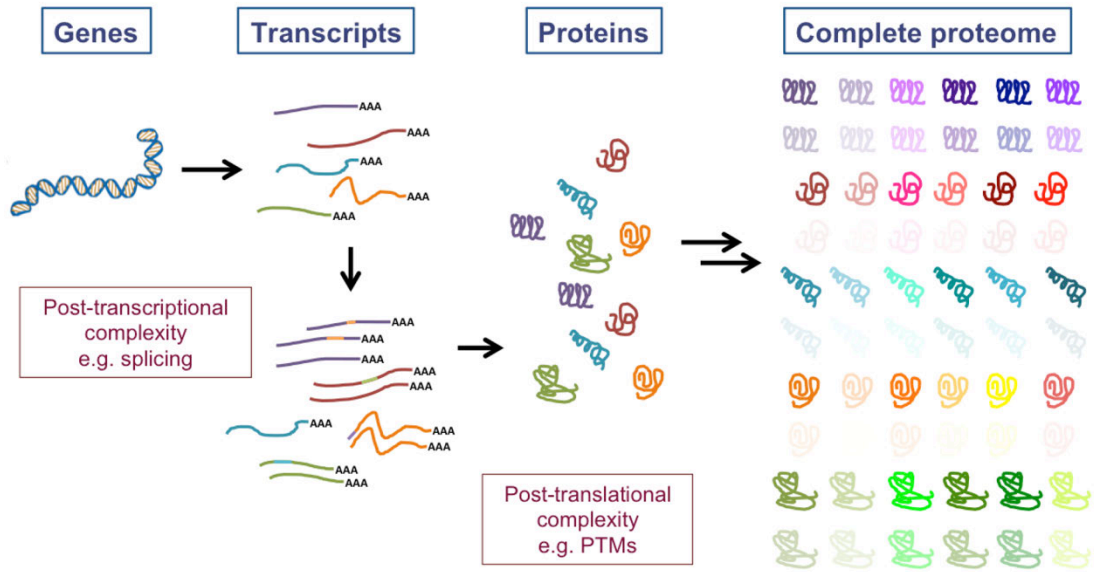


Figure 1.1: From genes to the complete proteome. Schematic depicting ways that diversity and complexity of the proteome can be increased by post-transcriptional and post-translational mechanisms.

1.3 PTMs and signalling

The variety and ubiquity of PTMs across the proteome has emerged as improvements in mass spectrometry have simplified their detection. Most amino acid residues can be modified in multiple ways (dependent in part on the flexibility of their chemistry) and multiple PTMs are more abundant than coded amino acid residues[13]. In tandem, a huge number of enzymatic reactions have been shown to catalyse these modifications, which creates a highly diverse and flexible set of products[13]. In charged, high abundance

proteins with amenable tryptic digestion fragments (classically histone proteins), PTMs have been relatively simpler to detect; the exquisite decoration of histone tails has been shown to provide an intriguing basis for gene regulation[14-16]. PTMs acting on the same amino acid can be competitive and antagonistic, neighbouring marks can occur with increased frequency or be mutually exclusively, and PTMs taken together therefore result in a huge regulatory framework[15,16].

As these chemical modifications can be introduced or removed dynamically on fully folded and active proteins, PTMs thereby provide an obvious conceptual mechanism for fine-tuned regulation and cellular control in real time[13]. PTMs, for example, regulate protein stability and rapid and controlled degradation, protein sub cellular localization (trafficking, nuclear/cytoplasm shuttling), control of binding interactions (recruitment or disruption of a interaction partner), control of enzyme activity (on/off, up/down), control of a protein substrate, and structural integrity (such as the construction of collagen fibrils or myelin development)[13,17-22]. It is therefore hardly surprising that much of the temporal and spatial information required for cell signalling is transmitted through PTMs.

In terms of the biological significance of these effects, complicated and abstract environmental signals are commonly transmitted through PTM cascades. The end point may be the stability or activity of a genetic effector that has consequent influence on gene expression. Some classic examples in the literature are metabolic/energy sensing (e.g. the mTOR system which senses amino acid availability)[23], low oxygen sensing (e.g. the Hypoxia Inducible Factor (HIF) system which is mediated through oxygen-dependent hydroxylations and the proteasome)[24], nitric oxide gas signalling (the NOS system)[25], heat-independent mechanisms to detect time (e.g. circadian rhythms mediated through phosphorylation feedback loops)[26-28], mechanical cues and stresses (e.g. the YAP/TAZ signalling pathway)[29], or spatially

resolved signals across and between different cells (e.g. the Wnt signalling pathway)[30].

1.3.1 PTMs and signal transduction: the Wnt example

Of these, perhaps the most classic effect of PTMs is in orchestrating signalling cascades (best studied, but by no means exclusive to phosphorylation)[31]. One PTM increases (or decreases) the activity of enzyme A; this enzyme A then modifies a second enzyme B affecting its activity. This can continue in complicated cascades of enzymes which might for example carry signals from the outside of a cell to the nucleus, rapidly transmit or amplify signals, or allow interconnectivity between different signalling pathways. Another classic effect of PTMs is in regulating protein stability (best characterized, but by no means exclusive to ubiquitination)[32,33]. Individual proteins that are tagged with the ubiquitin mark (through Lys48-linked chains) can be targeted for rapid and highly selective degradation by increased affinity to components of the proteasome[32,33]. The two systems (ubiquitination and phosphorylation) show widespread interconnectivity[34]. An illustrative example is the simplified picture of the mechanism of degradation of beta-catenin (β-catenin) in the Wnt signalling pathway[30,35,36]. The E3 ubiquitin ligase β-TrCP1 can utilize the transcription factor β-catenin as a substrate for ubiquitination[30,35,36]. This is mediated by interacting with a short N-terminal motif on β-catenin if it has been multiply phosphorylated[30,35,36]. This short motif is phosphorylated by two different kinases, firstly by the priming kinase Casein kinase 1 (CK1) and secondly by Glycogen synthase kinase 3 (GSK3)[30,35,36]. GSK3 only uses the singly phosphorylated N-terminal motif as a substrate, catalyzing the second and third phosphorylations[30,35,36]. If left unphosphorylated, however, β-catenin is stabilized, translocates to the nucleus and drives a transcriptional programme[30,35,36]. The requirement for two kinases integrates a specific signal between the highly promiscuous and constitutively active GSK3 and a context dependent and more specific second kinase[30,35,36]. β-catenin is rapidly turned-over at resting state, but if a signal disrupts the phosphorylation (or interrupts recognition by the

proteasome), then the transcriptional effects of β-catenin become switched on[30,35,36]. This is a typical cell-signalling paradigm (a signal is transmitted by temporarily pausing the continuous degradation of a transcription factor, in this case by disrupting a constitutive enzymatic activity). Other similar signalling paradigms have been observed throughout biology, such as the HIF system used in detecting hypoxia[24]. This discovery won the Nobel Prize in Physiology or Medicine in 2019 – the press announcement (7th October 2019) was made in the week this thesis was submitted.

1.4 Citrullination

Citrullination is a PTM made to peptidyl arginine, where the side chain primary ketimine (=N-H) is replaced with a ketone (=O) functional group[37]. The replacement of the nitrogen with oxygen results in a small mass difference (0.98 Da) and small change in sterics, but reverts both the positive electrostatic charge of arginine to the overall neutral (and *∂*- negative) charge of citrulline as well as two of arginine's five hydrogen bond donor sites to acceptors (Figure 1.2)[37].

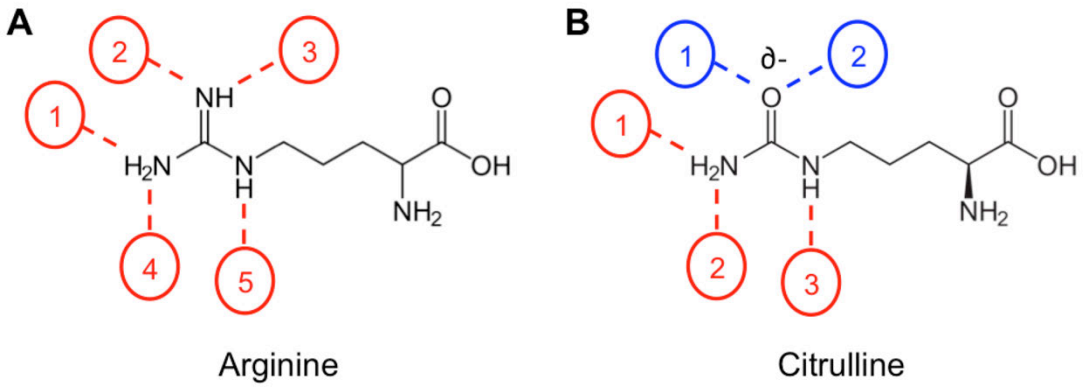


Figure 1.2: Comparison of the chemical structure of L-arginine with L-citrulline. Hydrogen bond donor sites are numbered and labelled in red with hydrogen bond acceptor sites in blue.

The free amino acid *L*-citrulline was first isolated in 1914 from watermelon juice[38], with precedent in nature in several enzymatic processes. Firstly, it is a product of deimination by free arginine deiminases (ADI) as part of arginine

catabolism and in arginine and proline metabolism (this biosynthetic pathway is distributed in some, but not all, bacteria)[39]. Secondly it is produced by ornithine transcarbamylase (OTC) as an intermediate product in the urea cycle, where it is subsequently converted back to arginine in two steps by argininosuccinate synthase (ASS) and argininosuccinate lyase (ASL)[40,41]. Thirdly citrulline is a byproduct of nitric oxide synthases in nitric oxide signalling[42-44]. Lastly it is generated by dimethylarginine dimethylaminohydrolase (DDAH), that removes dimethyl arginine from degraded methylated proteins; free dimethylarginine otherwise inhibits nitric oxide synthases in nitric oxide signalling[45,46]. These processes are summarized in Figures 1.3 and 1.4.

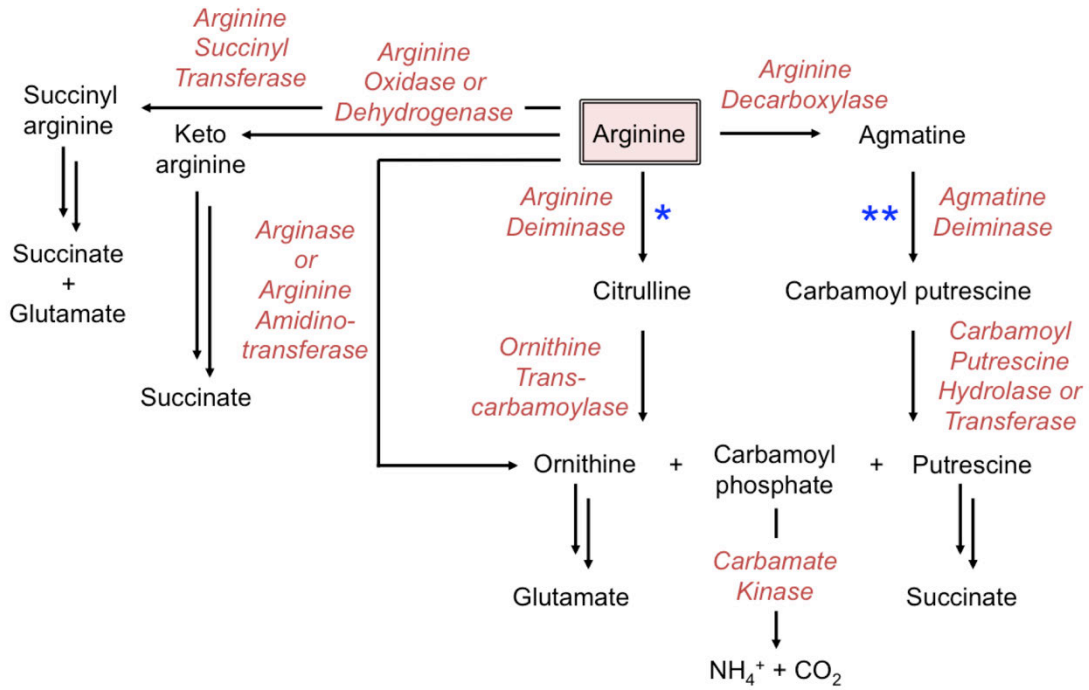


Figure 1.3: Overview of arginine catabolic pathways in bacteria. The first few steps of six major arginine catabolic pathways that are found in bacteria are shown. Intermediates are labelled in black and the corresponding enzymes labelled in red italics above the single reaction arrows in black. Two black reaction arrows denote that several enzymatic steps lead to the eventual product of the pathway, which are abbreviated for clarity. The reactions catalysed by Arginine deiminase (ADI) and by Agmatine deiminase (AgD) enzymes are labelled with one or two blue asterisks respectively.

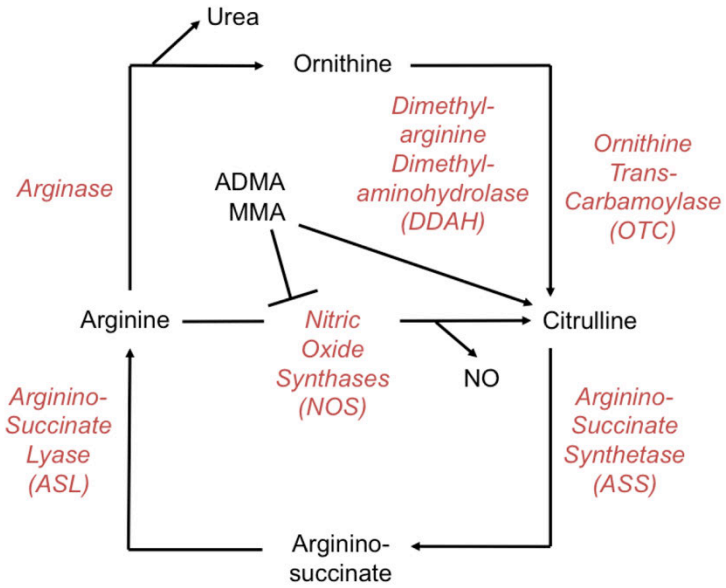


Figure 1.4: Overview of the urea cycle with a focus on citrulline. The urea cycle and related reactions are shown with reactions denoted by single black arrows. The corresponding enzymes are labelled in red italics beside each reaction. Symmetric dimethyl arginine (ADMA) and monomethyl arginine amino acids inhibit the activity of nitric oxide synthases, but are converted to citrulline by *Dimethyl-arginine Dimethyl-aminohydrolase (DDAH).*

Citrulline was first found in a protein context as early as 1930 when the chemical structure of the free amino acid was also assigned[47-49]. Citrulline was subsequently identified by *Rogers et al.* in hair follicle protein hydrolysates where it was found to be highly abundant[50,51] and in 1971, it was revealed to be a component of myelin by Moscarello and Wood[52]. In 1977, the enzyme responsible for the citrulline content in hair follicles was partially purified and the calcium-dependent post translational enzymatic conversion of arginine to citrulline was first clearly demonstrated *in vitro*[53]. In 1981, the citrullinating enzyme was fully purified from rats, further characterized, and named peptidyl arginine deiminase (PADI or PAD)[54]. Since then, several studies have revealed there to be 5 PADI paralogues in humans (named PADI1, PADI2, PADI3, PADI4 and PADI6 – PADI5 was shown to be the same as PADI4) and many roles for citrullination have been identified in different tissue contexts[55]. PADIs are reported to be distributed only in higher eukaryotes, and are widely distributed only in vertebrates; fish

possess a single PADI, with five paralogues in mammals[55-57]. This is discussed further in the short introduction to Chapter 3.

The process of citrullination refers specifically to the post-translational conversion of arginine to citrulline in protein substrates (Figure 1.5). Predominantly, citrullination has been described as the catalytic activity of the PADIs, but has two other known precedents in the literature. In the first, the reaction is catalyzed by a modified free agmatine deiminase (agmatine is decarboxylated arginine) identified in the bacteria responsible for periodontitis *Porphyrmonas gingivalis* (referred to throughout as pPAD) (Figure 1.3)[58]. The second is catalyzed by a modified free arginine deiminase identified in *Giardia lamblia* (referred to throughout as gADI) where it has a role in its primitive immune system[59] (Figure 1.3). These other enzymes (pPAD and gADI) show different substrate preferences to PADIs, are not calcium dependent, and possess very divergent protein sequences[59-64]. A "decitrullination" reaction, the reverse catalytic process of citrullination, has not been identified to date.

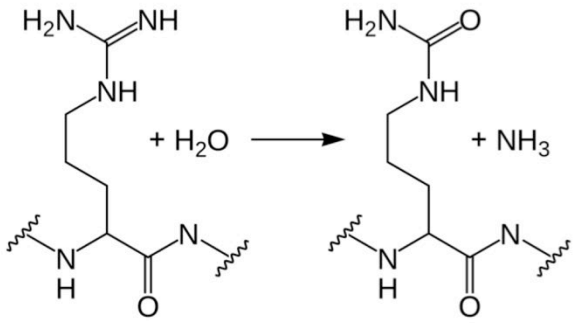


Figure 1.5 Citrullination reaction catalyzed by the PADI enzymes. PADIs catalyze the conversion of peptidyl arginine residues into peptidyl citrulline using water with equimolar release of ammonia. Calcium ions are required for the formation of the active conformation of PADIs by allosteric binding to the enzyme, but do not take part directly in catalysis.

All citrullinating enzymes (and several other enzymes) make use of the same pentein protein fold[61,65-68] (Figure 1.6). This is highly divergent in terms of amino acid similarity among members, but the active site triad (Cys- His- Asp) is shared between pentein-fold containing enzymes. Pentein-fold

containing enzymes catalyse hydroxylation, dihydroxylase and aminohydrolyse reactions and have a wide range of substrates. Although most members show a preference for free amino acids, PADIs, the extended *Giardia lamblia* gADI, and *Porphyrmonas gingivalis* pPAD show tolerance for protein substrates[58,59]. PADIs, however, exclusively convert protein substrates[69].

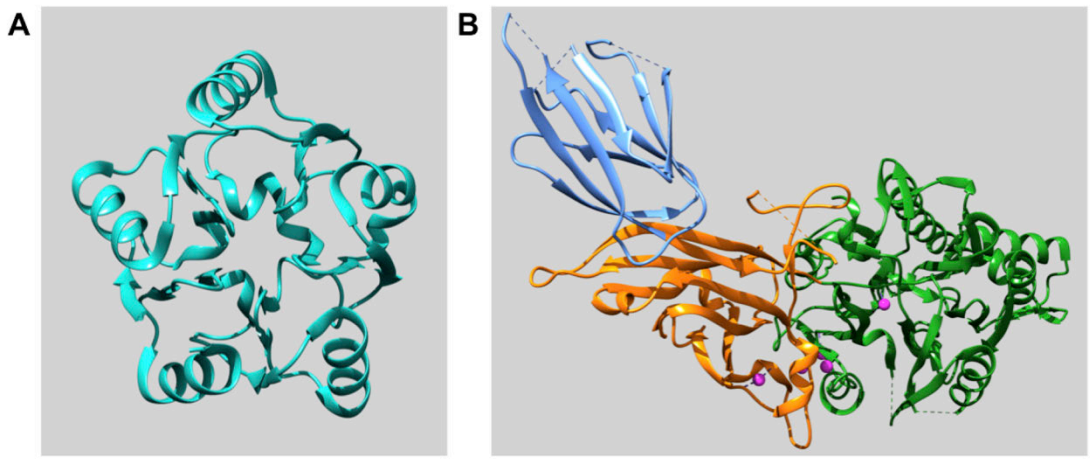


Figure 1.6 The catalytic PAD_C domain possesses a pentein fold: A: Crystal structure of ribosome anti-association factor eIF6 (1G61, cyan) showing the pentein fold which comprises five αββαβ subdomains around a fivefold axis of pseudosymmetry. **B:** Crystal structure of human PADI4 (1WD9) shows the three domains PAD_N (cornflower blue), PAD_M (orange), PAD_C (forest green) in a view which rotates the structure to show the same orientation of the pentein fold contained in the catalytic PAD_C domain showing the same pentagonal fivefold pseudosymmetry and subdomain structure but with additional conserved inserted regions. Calcium ions are shown as magenta spheres.

PADI enzymatic activity is stringently regulated[70,71]. Allosteric binding of up to six calcium ions structures the active site for catalysis (Figure 1.7, Figure 4.1)[70,71]. This is discussed in more detail in the introduction to Chapter 4. The proposed mechanism for the PADIs makes use of a nucleophilic cysteine to form a tetrahedral S-alkylthiouronium adduct intermediate[72-76]. The intermediate collapses and ammonia is released. Subsequently, water, activated by the core His residue, then attacks as a nucleophile and regenerates the tetrahedral S-alkylthiouronium intermediate. This time, the core cysteine thiolate is eliminated to leave peptidyl citrulline. Interestingly

while PADI1-3 and PADI4 are thought to use a reverse protonation mechanism (with a narrow optimal pH window that can support a deprotonated core cysteine, and a protonated core histidine), PADI2 is instead proposed to use a substrate-assisted mechanism (where the positively charged guanidinium substrate improves the cysteine nucleophile by lowering its pKa). These conclusions are supported by various pH-dependent inactivation, kinetic, mutagenesis, pH rate profile, solvent isotope effect, and solvent viscosity effect studies[72-76].

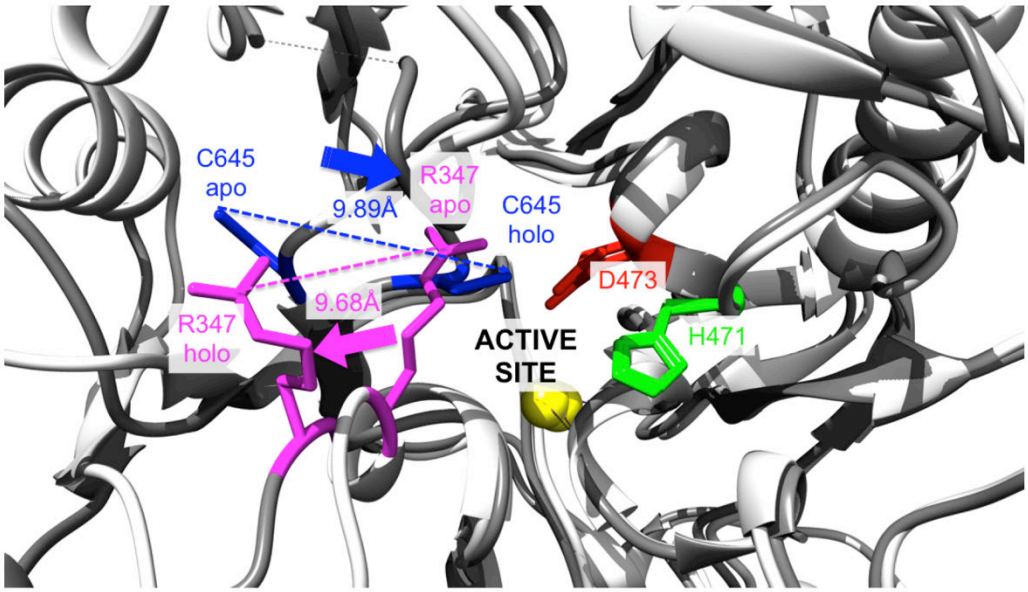


Figure 1.7: Structure of the PADI active site with and without allosteric calcium binding. The structure of PADIs in low calcium is called the apo-form, whereas the fully calcium coordinated enzyme conformation is called the holo-form. The crystal structure of the active conformation of PADI2 is shown in dark grey (4N2C) and the inactive conformation is shown in white (4N20). These two structures were superposed in Chimera using MatchMaker to align the structures with the Needleman-Wunsch algorithm according to the BLOSUM-62 matrix. The figure denotes the apo-form of PADI2 (light grey) superimposed on the holo-form of PADI2 (dark grey) showing the active site. The nucleophilic cysteine C645 is shown in blue which moves 9.89 Å into the active site in the active holo enzyme conformer. Conversely a "gatekeeper" arginine R347 (magenta) located within the active site in the apo (inactive) conformer leaves the active site in the active conformation to allow substrate binding. The other two residues in the catalytic triad (D473, red and H471, green) do not differ greatly in their position in the active site between inactive and active enzyme conformations.

1.5 Properties of arginine

As an amino acid residue, the intrinsic chemical properties of arginine make its modifications particularly biologically interesting. Arginine is protonated and positively charged at physiological pH (as such it is a poor nucleophile). It possesses five hydrogen bond donors that can be used to interact with polar groups. This is characterized by the formation of arginine salt bridges with phosphate groups (found for example in nucleic acids or phosphoproteins) which are twice as strong as those formed by lysine[37]. In addition, arginine has a long and flexible side chain, which aids interactions. It is therefore unsurprising that arginines are one of the key amino acid residues for mediating protein interactions to negatively charged nucleic acids and RNA and are over-represented at protein interfaces (unlike lysine, aspartate or glutamate)[77-80]. The abundance of arginine modifications in histones is also of considerable importance. More specialized consequences of modifications to arginine derive from specific contexts such as their abundance in nuclear localization signals[81], in serine protease inhibitors[82], or in liquid-liquid phase separation through interactions with tyrosine[83].

1.6 Specific cell biological consequences of citrullination

Many direct effects of citrullination on proteins have been identified in line with the functions expected for modifications to arginine. This includes regulating protein and nucleic acid interactions[9,84,85], altering sub-cellular localization[86-88], modifying properties of protein structure[9,89-91], and most recently in modulating liquid-liquid phase separations[83,91]. A handful of specific examples of more specialized modifications of arginine are discussed below with a focus on citrullination (Figure 1.8).

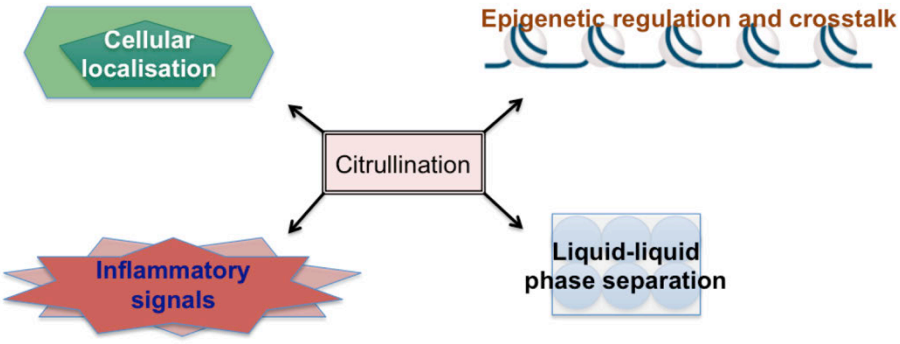


Figure 1.8: Some specific cell biological consequences of citrullination. A schematic depicting some of the cell biological consequences of citrullination that are focused on in more detail below.

1.6.1 Epigenetic regulation and crosstalk with arginine methylation

PTMs made to arginine that have been discovered to date, in addition to citrullination, include three distinct methylation marks: monomethylation, asymmetric dimethylation and symmetric dimethylation, as well as less canonical modifications such as arginine phosphorylation, ADP-ribosylation, methylglyoxal adduct formation and recently C-5 hydroxylation[92-96].

The prevalence of arginine methylation appears to be comparable to that of phosphorylation/ubiquitination, but only nine PRMTs have been identified (as compared with hundreds of identified kinases/ubiquitin ligases and approximately 40 lysine methyltransferases)[97,98]. Possessing both forward and reverse reactions has elevated the importance of certain PTMs such as lysine acetylation and methylation in the literature as it is assuming this more clearly implies a dynamic role. As such, the first call to fame for citrullination was over the hypothesis that it might either act as an arginine demethylase (or at least antagonize arginine methylation) and might therefore be able to elevate arginine methylation to the same level. This hypothesis for citrullination was strengthened by the realization that methylated arginine residues overlap with citrullination sites: arginines 2, 17, and 26 in histone H3 are substrates for methylation by CARM1[99-102] and arginine 3 in histone H4 is methylated by PRMT1[7,103-105]. Indeed, citrullination is found to be antagonistic to arginine methylation and many examples of epigenetic

crosstalk have been identified between these marks and the importance of antagonistic marks (such as citrullination) are of clear regulatory importance (for example *Sharma et al*. for Pol II)[106].

However, although the citrullination mark has been shown to be antagonistic to methylation[107], PADIs do not appear to act directly on methylated arginine in physiological contexts (and in any case do not recover arginine) and show clear independent roles from merely acting in opposition to arginine methylation[103,108]. There is continued interest in the possibility of a true arginine demethylase. Although some candidate arginine demethylases show activity on peptides in vitro, biological significance has yet to be demonstrated[109,110]. Additionally, there is also capacity for nucleosome turnover to account for a dynamic quality to arginine methylation– in the absence of a decitrullinase or arginine demethylase dynamic methylation and citrullination states could still readily occur[111]. The caveat of course is that lysine methylation was also once thought to be irreversible and its dynamic nature rationalized similarly until the discovery of LSD1/KDM1A[112-114].

1.6.2 Cellular localisation

GSK3β is an important cellular kinase with pools in various cellular compartments[115]. Complicated nuclear import mechanisms appear to regulate its cellular location along with possible mechanisms to retain a cytoplasmic pool[115]. An interesting study identified a classical monopartite type nuclear localization signal (NLS) in GSK3β where repeated basic (lysine and arginine) residues are located[87]. Accordingly, Δ9-GSK3β-HA was found to be at 40% lower levels than wild-type GSK3β-HA in the nucleus)[87,116]. Subsequently PADI4 was found to citrullinate the N-terminus of GSK3β directly (R3, R5) increasing its nuclear localization (R->K point mutants were no longer regulated)[87]. PADI4 depletion reduced nuclear localization of GSK3β; stable expression of PADI4 increased nuclear GSK3β, whereas catalytically inactive PADI4 showed no effect[87]. In a recent study, the protein complex mTORC1 was also found to regulate the nuclear localization of

GSK3β, presumably through a separate mechanism as it similarly affected the GSK3α isozyme that has a truncated N-terminus[116]. In cells grown in serum, GSK3β is increasingly retained in the cytoplasm, but in serum withdrawal conditions but the nuclear level of endogenous GSK3 increased[116]. These observations have not been fully explained.

In a different mechanism, citrullination affected the nuclear localization of the p65 subunit of nuclear factor κB (NF-κB p65) by increasing its interaction with a nuclear importin[88]. PADI4 was shown to interact with and citrullinate four arginines in the N-terminal RelA homology domain of NF-κB p65 in neutrophils[88]. This caused increased binding of importin a3 to p65 (with no effect observed to R->K point mutants of p65)[88]. This was correlated with increased nuclear import of p65, and in turn drove increased expression of inflammatory cytokines IL-1β and tumour necrosis factor alpha (TNFα)[88]. PADI4 was thereby shown to affect the transcriptional activity of nuclear factor κB (NF-κB) driven by lipopolysaccharides (LPS)[88]. In another example, citrullination of nucleophosmin (NPM1) at Arg197 by PADI4 was shown to shift its subcellular localization from the nucleoli to the nucleoplasm[86]. In a final example, citrullination of the bipartite nuclear localization signal of Inhibitor Of Growth Family Member 4 (ING4) did not affect its nuclear localization directly[85,117]. ING4 had been shown to induce acetylation of K382 of p53, increasing p53 transcriptional activity and thereby driving p21 expression, mediated through binding to p53 at the NLS region[85,117]. Citrullination of ING4's NLS by PADI4 disrupted the interaction with p53, decreased the acetylation of p53 and interfered with downstream p21 expression[85,117]. This also caused more rapid degradation of ING4[85,117]. Although NLS type motifs make good candidates for direct effects of regulation by citrullination due to their arginine content, they do not follow a single well-established mechanism of action. In the instance of ING4, the bipartite NLS motifs appear to primarily make use of lysine residues, which may explain the differing behavior of citrullination in this example.

1.6.3 Extracellular effects and inflammatory signals

Various inflammatory signalling proteins are modulated by citrullination. Host defence peptides are an ancient part of mammalian innate immune systems: one such peptide, LL-37 in humans, binds to the negatively charged lipid domains of endotoxin, to LPS, to lipoteichoic acid (LTA) and polyI:C and also to cell free immunostimulatory DNA[118,119]. LL-37 acts to dampen the inflammatory response to these agents (such as the release of TNFα and nitric oxide by macrophages)[118,119]. The cationic nature caused by multiple arginine and also lysine residues within LL-37 aids the interaction with these negatively charged inflammatory agents[118,119]. LL-37 becomes citrullinated in conditions of sepsis and the citrullinated peptide binds very weakly to these agents, which abrogates the reduction in proinflammatory effect[118,119]. Inflammatory effects of citrullination have also been found to modulate the activity of chemokines including IL-8 (CXCL8), CXCL10, CXCL11, and CXCL12[120,121]. In these cases, by contrast to LL-37, citrullination acted to dampen signalling potency[120,121]. Citrullination of CXCL10 and 11 reduced chemoattraction and signaling capacity, without affecting receptor binding[120,121]. Similarly, citrullination of IL-8 reduced the induction of neutrophil extravasation and chemotaxis[120,121]. Given that 14% of natural leukocyte-derived IL-8 was found to be citrullinated and the efficiency of PADI catalysis in serum, rapid innate immune modulation effects may be found in other NETotic conditions (cells undergoing cell death via NET formation) and in various immune contexts[120,121]. Finally, a recent proteomics screen of the rheumatoid arthritis (RA) associated citrullinome identified several citrullinated serine protease inhibitors (P1-Arg-containing serpins) from RA serum, synovial fluid, and synovial tissue, confirming earlier work in vitro and in RA patients[122,123]. The serpin-protease complex involves an interaction between a protease that attempts to catalyse proteolysis of the inhibitory serpin[122,123]. This takes place on a reactive peptide bond between arginine and serine contained in the serpin amino acid sequence, which becomes bound with high affinity and irreversibly inhibits the protease stoichiometrically[122,123]. Citrullination of this arginine, found in antiplasmin,

antithrombin, t-PAI, and C1 inhibitor, was shown to disrupt interaction with the corresponding proteases and prevented inhibition[122,124]. Extracellular serpins (which comprise about two-thirds of all human serpins) have been shown to modulate proteolytic cascades found in blood clotting (antithrombin), in inflammatory and immune responses (antitrypsin, antichymotrypsin, and C1-inhibitor) and in tissue remodelling (PAI-1)[124-126]. Given the inhibitory activity of serpins on neutrophil elastase, granzyme B, cathepsins and others, further work to elucidate the crosstalk between neutrophils, NETosis and the activation of PADIs in these extracellular immune networks would be very interesting. These may also be relevant to the effect of PADI enzymes in cancer contexts[127,128].

1.6.4 Liquid-liquid phase separation

Another interesting effect that can be modulated by post-translational modifications to arginine are liquid-liquid phase separation type interactions[83]. Cells have different compartments for isolating complex biochemical reactions in space, including well known membrane bound compartments such as the mitochondrion or lysosome. Other compartments, however, are not membrane bound such as nucleoli, centrosomes, Cajal bodies, stress granules and P granules. In these cases, it has begun to emerge that these unbound compartments can arise from the phase separation of complex biological mixtures into liquid droplets[129,130]. If the components of the compartment have a higher affinity with each other than the surrounding mixture they can demix and form phase separated liquid droplets[131]. For the fused in sarcoma (FUS) RNA binding protein, comprised of a low complexity disorded domain (LC domain) attached to a structured C-terminal domain comprised of multiple RGG motif repeats, its solution can reversibly revert between a dispersed state into liquid droplet and hydrogel states[83]. These proteins have a role in producing ribonuclear protein granules. Recent experiments have revealed that FUS phase separation can be modulated by the arginine methylation state of RGG motifs in the structured C-terminal domain[83]. These RGG arginines, predominantly in a

methylated state after conversion by PRMTs, interact with tyrosines in the N-terminal domain via cation pi interactions[83]. Unmethylated arginine residues have the strongest interaction, methylated arginines somewhat weaker, but citrullinated residues abolish this interaction[83]. Adding a small amount of unmethylated arginine FUS to a predominantly methylated arginine FUS pool drives rapid phase separation[83]. This is particularly interesting for citrullination, as RGG motifs have separately been identified as a common PADI substrate motif (comprising approximately 1/5 of PADI substrates)[91]. Citrullination of various RGG motif proteins, including FUS, were shown to abolish their protein aggregation[91], likely both by competitively eliminating arginine methylation but also through eliminating cation-π interactions with tyrosine residues[83]. The methylated RGG motif is also well known for mediating an interaction with the Tudor domain containing protein SMN (an arginine methylation reader protein)[132]. This interaction was shown to be suppressed by PADI4 citrullination in the same context[91]. Precise mixtures of modified and unmodified FUS proteins provide a possible physiological mechanism to interconvert between phase separation states. This is especially interesting in light of the many ribonuclear proteins containing RGG motifs and low complexity domains and also in light of roles for phase separation emerging in other nuclear process such as in interphase chromosome structure[133] or heterochromatin domain compartmentalization[134,135]. It is particularly interesting in light of the large number of ribonuclear proteins, RNA binding proteins and RGG motif containing proteins that are substrates for citrullination (personal communication, unpublished tandem mass spectrometry (MS/MS) data from *Christophorou et al*.). As such, other roles for mechanisms of PTM regulation in this area are likely to emerge.

Along the same lines, a recently identified citrullination of Arg1810 in the C-terminal domain of RNA polymerase II (RNAP2), also a target of arginine methylation may also be connected[106]. The C-terminal domain (CTD) of RNAP2 is dynamically modified, which recruits different protein complexes

that modulate transcription[136,137]. Loss of citrullination of RNAP2 appeared to result in the accrual of RNAP2 close to transcriptional start sites, whereas citrullinated RNAP2 recruits the positive transcription elongation factor P-TEFb that releases paused RNAP2 and promotes gene expression[106]. Given that RNA binding proteins such as FUS can sequester the CTD of RNAP2 in phase separated droplets[138,139], this may offer a possible extension to the relevance of modified arginine in liquid-liquid phase separation and transcription– balanced between methylated arginine on the one hand and citrullinated arginine on the other[140-142].

1.7 Peptidyl arginine deiminases (PADIs)

There are five PADI paralogues in mammals (PADI1-4 and PADI6), which are carefully regulated, both transcriptionally and enzymatically (Figure 1.7 and 1.9)[69,143]. The genomic context of the five mammalian paralogous genes is conserved across species. PADI1, PADI3, PADI4, and PADI6 are located in close proximity and transcribed in one direction, whereas PADI2 is located on the reverse strand and at a greater distance to the cluster of paralogues (Figure 1.10)[143]. PADI paralogues show restricted expression to specific tissues and cell types and also show stringent restriction on enzymatic activity, dependent on binding calcium ions (Ca^{2+}) and with some variation in substrate selectivity[55,124,144-146].

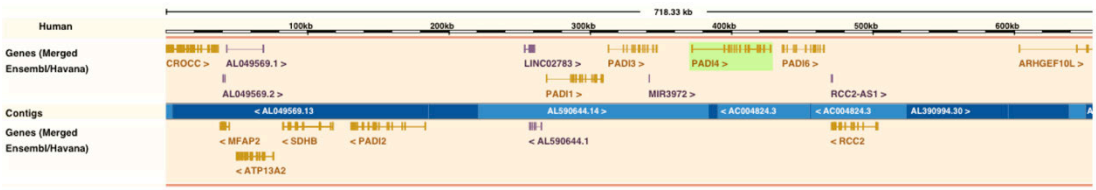


Figure 1.9: Genomic region containing the five PADI genes. A: The genomic region in humans containing the five paralogous PADI genes is shown as a track using Ensembl with PADI4 highlighted in green. PADI2 is located further from the other 4 paralogues in the reverse strand direction.

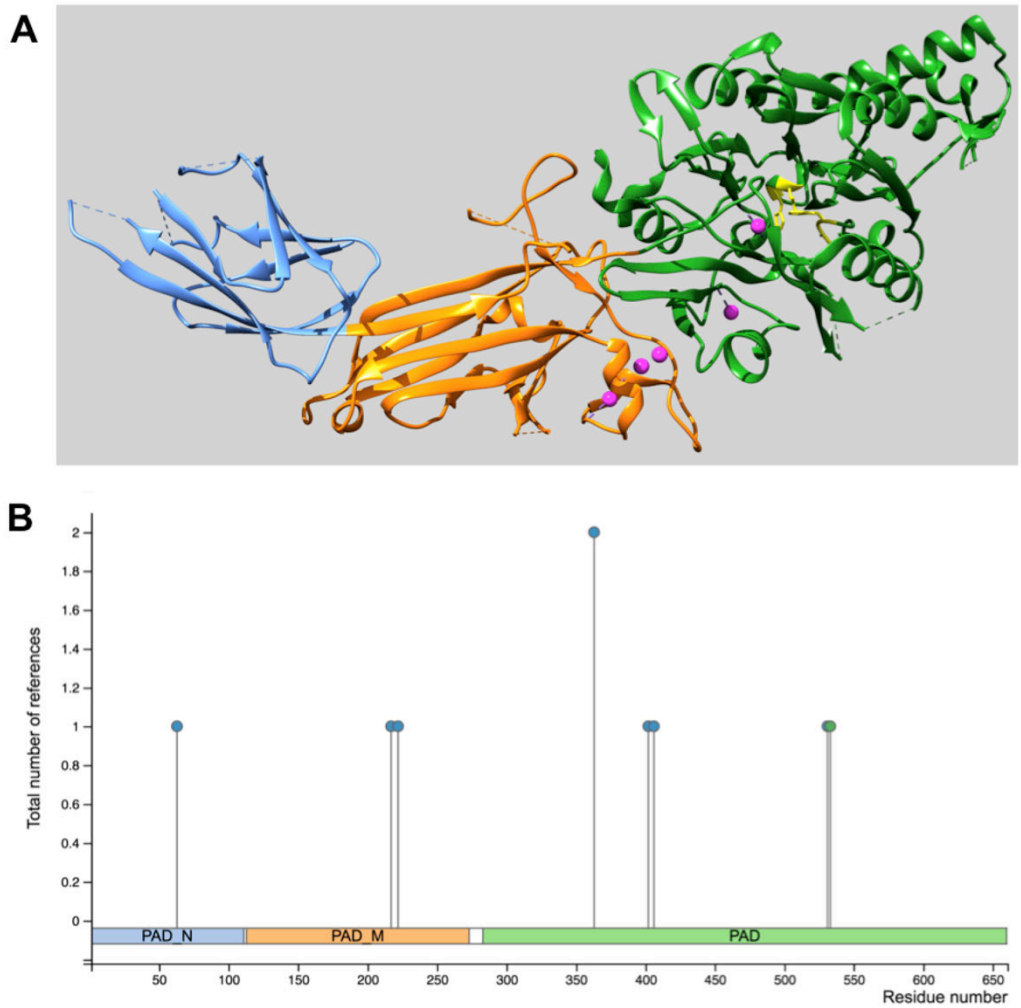


Figure 1.10: PADI4 and known post-translational modifications: A: Crystal structure of calcium soaked human PADI4 (1WD9) shows the three domains PAD_N (cornflower blue), PAD_M (orange), PAD_C (forest green), with calcium ions shown as magenta spheres and residues in the catalytic triad shown in yellow. **B:** Domain cartoon of PADI4 shows the currently annotated phosphorylation sites in PhosphoSite detected in either high-throughput or low-throughput proteomic experiments in the literature.

1.7.1 Physiological roles for PADI paralogues

In terms of normal biological functions, PADIs are relatively poorly characterized, especially compared to their roles in disease contexts[69]. PADI1 and PADI3 are known for a handful of roles in hair and epidermal contexts[69]. Terminally differentiated epithelial cells undergo controlled cell death with broad cellular proteolysis: citrullination of filaggrin results in the dissociation from keratin filaments which aids hydrolysis[147]. By contrast, citrullination of trichohyalin increases its protein solubility, aiding

transglutaminase mediated cross-linking and strengthens the hair follicle[148,149]. Citrullination of vimentin, alongside phosphorylation, is proposed to play a role in the regulated polymerization and depolymerisation of vimentin proteins[150-153]. The role for PADI3 in keratinocytes is supported by genome wide association studies (GWAS) that find SNPs in PADI3 that are significantly associated with uncombable hair syndrome, hair shape, and hair colour[154-156].

PADI6 has the clearest biological phenotype, but is enigmatic as its catalytic activity has yet to be demonstrated. PADI6 is essential for early cleavage divisions: PADI6 knockout mice are infertile, with fertilized oocytes unable to progress past two cell division and form oocyte cytoplasmic lattices[157-159]. This phenotype has recently been confirmed in mutations in humans as well[160-162]. However, PADI6 shows some divergence in the conservation of the active site and calcium binding residues from the other paralogues (Figure 1.11). As the catalytic cysteine is replaced by alanine and two of the six binding residues in Ca1 and Ca2 are conserved differently, this has been widely taken as evidence that PADI6 cannot support enzymatic activity. However, the cysteine to alanine residue is sandwiched between two other highly conserved cysteines, which could provide alternative catalytic nucleophiles and it may be that PADI6 is merely regulated differently from the other paralogues enabling activity in vivo (Figure 1.11). Interestingly, PADI6 was found to undergo cell-cycle dependent phosphorylation, which allowed subsequent interaction with the 14-3-3 protein YWHAB[163]. This YWHAB interaction was confirmed structurally[164]. If catalytic activity by PADI6 can be clearly demonstrated, it is an intriguing possibility that a phosphorylation dependent interaction might provide a mechanism for PADI6 regulation.

PADI2 and PADI4 are the best-studied paralogues, but individual knockout mice are superficially healthy. Recent literature points to a role for PADI2 in oligodendrocyte differentiation[165]. PADI2 primarily, but also PADI4, are

known to extensively modify MBP in normal myelin, but have been shown to destabilize myelin biochemically and show up regulation in Multiple Sclerosis lesions[89,166-168]. It is not fully clear what roles PADI2 or PADI4 play in the normal functioning of myelin development. The other biological role for PADI2 and PADI4 is in innate immune defence. As early as the 1880-90s, Elie Metchnikoff discovered the role of neutrophils in undergoing phagocytosis of infectious agents[169]; it is particularly remarkable therefore that in 2004 a completely new mechanism of neutrophil response to infection was discovered: neutrophil extracellular trap (NET) formation[170]. Although PADI4 was assigned a role in neutrophils as having an essential role in NET formation[171,172], further characterization and more precise means to detect NET formation have meant the precise roles of PADI4 and PADI2 are less clear and conflicting data exist from mouse models[173]. Different stimuli (some involving citrullination, some not) can induce NET-like structures and different types of NET-like structure formation are induced by different stimuli[174,175]; it is not fully clear which forms are physiological or relevant as a response to infection. Further work will also be required to tease out the precise role of citrullination. This is both because highly pleiotropic stimuli are commonly used as NET agonists[174-177] and no tools or stimuli to precisely activate PADI enzymes exist. PADI4 does appear to be responsible for the citrullination observed in neutrophils[173,177], but conclusions beyond that should be more tentative based on the present data.

The final role for PADIs has emerged in development. Pan-PADI inhibition halts embryonic development at the 4-cell stage, and PADI4 was found to play a role in preimplantation development by use of RNA interference[9,157]. PADI4 was also shown to be induced and inference of its enzymatic activity switched on from citrullination of histone 3 during reprogramming of mouse neural stem cells to induced pluripotent stem cells (iPS cells)[5,9,178]. There is an additional described role for PADI1 in the uterus/early embryo development[179]. Mechanistically, PADI4 affects chromatin decondensation as was shown in neutrophils by histone citrullination[8,180]. This effect was then

also found to be relevant in the context of reprogramming, where direct citrullination of linker histone H1 displaced it from chromatin in vitro[9]. Whether this is a precise enough mechanism to describe the fine-tuning of reprogramming efficiency is much less clear. This is especially true since the scale of decondensation in neutrophils is so dramatic that chromatin architecture is completely dismantled[181,182]. In reprogramming cells, however, the resultant iPS cells can redifferentiate down any lineage. Mechanisms for the regulation and activation of PADIs, especially if their activation has such a dramatic cellular effect, remain obscure.

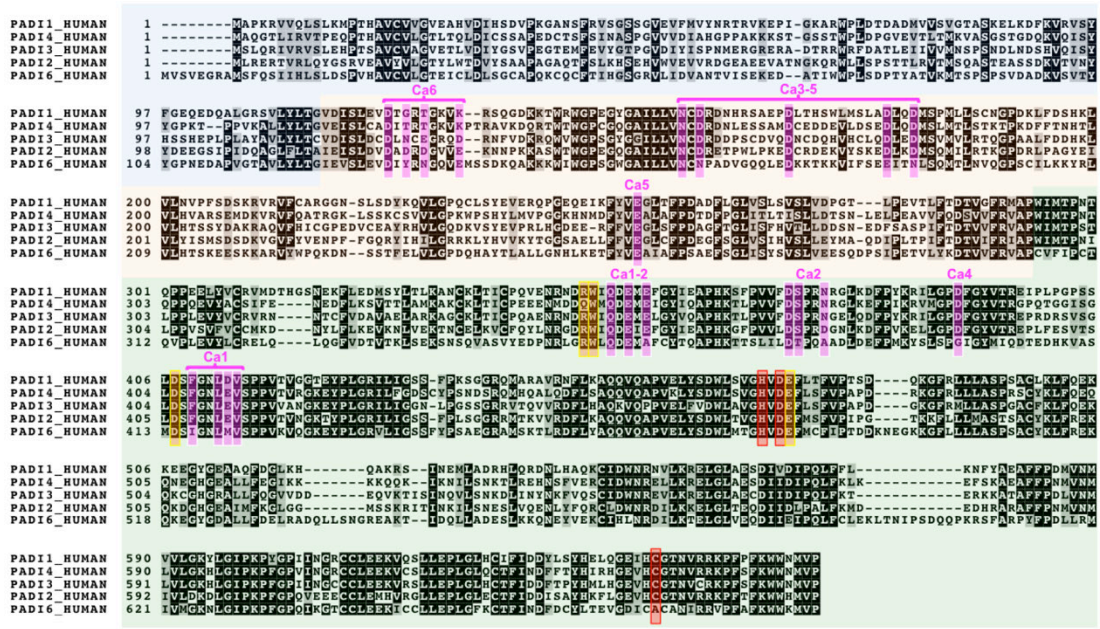


Figure 1.11: Multiple sequence alignment of all human PADIs. Human PADI1, PADI2, PADI3, PADI4 and PADI6 sequences were obtained from UniProtKB, aligned using Mafft G-ins-I and figure prepared using BoxShade with 70% consensus for greyscale shading. The catalytic triad residues are coloured in red, other active site residues coloured in yellowy orange, and residues used in PADI2 for calcium binding coloured in magenta, noting that calcium site 6 of PADI2 is not conserved for example in PADI4. PADI protein domains are shaded with PAD_N in cornflower blue, PAD_M in orange, and PAD_C in forest green.

1.8 PADIs and disease

As this chapter has established, although many specific examples have accumulated in the literature of the effects of citrullination in regulating cell biological events, clear physiological roles for PADIs are less well established. Much more work, however, has been undertaken on the

connections of aberrant citrullination to disease, predominantly resulting from too much citrullination[55,183]. Intriguingly, these connections have been to diverse autoimmune diseases, about which little is known in terms of cause or mechanistic biological understanding. In addition, roles for PADIs have been found in a variety of cancer contexts. Particularly strong connections have been found to multiple sclerosis (MS) and rheumatoid arthritis (RA). Other connections to a diverse range of other autoimmune and some non-autoimmune disease etiologies have also been discovered, including psoriasis[184,185], systemic lupus erythematosus[186-188], type I (autoimmune) diabetes[189-191] and Epstein barr virus; as well as in non-autoimmune contexts such as glaucoma[192], deep vein thrombosis[193], sepsis[194,195], allergy[196], prion diseases[197], Alzheimer's[198] and hypoxic injury[55,199,200].

The double knockout of PADI2 and PADI4 has not been produced, but single knockouts of PADI2 and PADI4 are generally viable and fertile[55]. Limited penetrance of preweaning lethality has been reported in some colonies. Knockout studies in mice have produced confusing and conflicting results in disease models. Where this has been rigorously tested for, compensation effects with other paralogues have been observed. In the overexpression model of PADI2, endogenous expression of PADI4 was found to be elevated at both the RNA and protein level[201]. Results in general point to the likelihood of overlapping functionality and auto-regulatory effects between different paralogues[171-173,201-206], but the absence of a double knockout means no firm conclusions can be drawn on compensation effects. Consistent with this, many studies using PADI inhibitors have shown potentially therapeutic effects of pan-PADI inhibition in disease models. As enzymes, PADIs generally are thought to pose a promising therapeutic target[183] and a possible avenue to a better understanding of disease mechanisms. The disease connections to MS, to RA and to cancers are reviewed more carefully below.

1.8.1 PADIs and Multiple Sclerosis

Citrullination has emerged as a hallmark of multiple sclerosis (MS). MS is an autoimmune demyelinating disease causing lesions that affect white matter in the brain. Myelin basic protein (MBP) is a multi arginine containing cationic protein that forms the main protein component of myelin and is the second highest abundance protein in the central nervous system. Myelin is considered to be one of the key distinguishing features of jawed verterbrates; protein components of myelin surround lipids, which together form insulation around axons. MBP was found to be citrullinated at multiple residues in children, but the extent of citrullinated MBP decreases in postnatal development (from ~45% down to ~20% in healthy adults as a proportion of arginines that are found to be converted to citrulline)[89,166]. Interestingly MBP citrullination was also found to be elevated in multiple sclerosis (to the same levels as in early development ~45%) as compared to healthy control adult tissue (~18%), and are dramatically elevated in an acute fulminating form of MS, the Marburg variant (~80-90%)[89,166]. Various studies showed that citrullinated MBP exhibits an increased rate of proteolysis and disrupted protein-lipid bilayer interactions which result in destabilization of the myelin sheath[207]. This has been hypothesized to expose the immune system to the MBP fragments that are found to be auto-antigens in the disease, promoting the autoimmune aetiology. It is interesting that an increase of citrullination in the normal appearing white matter (NAWH) around lesions can be detected in MS patients over white matter taken from controls. In white matter, myelin protein components GFAP, MBP and MOG are all substrates for citrullination[55].

Provision of an immunodominant epitope of MBP (“immunization of animals with spinal cord homogenates, individual myelin proteins, or by adoptive transfer of myelin-specific T cells”) drives a mouse model of MS called experimental autoimmune encephalomyelitis (EAE)[203]. The research on citrullination and MS was somewhat confounded when a paper reporting a PADI2 knockout mouse model showed that the mice formed healthy myelin

and could also develop EAE[203]. Subsequent studies established from the same knockout PADI2 mice taken from the same colony[205], showed that MBP citrullination could in fact still be detected in these mice in contrast to the original report and cast doubts on those original conclusions[204,205]. This revised understanding is consistent with transgenic mice overexpressing PADI2 that show a myelination phenotype and a recent PADI2 knockout model that showed PADI2 is required for proper oligodendrocyte myelination[208]. Paralogue compensation has been suggested as a possible reason for these discrepancies[208], for which PADI4 is the likely candidate, but no conclusions can be made on the evidence to date[204]. As PADI4 is also found to be elevated in the NAWH of MS patients and endogenous PADI4 expression was found to go up when PADI2 was overexpressed, the expression of different paralogues may be co-dependent and intricately regulated. In a different model, PADI2 overexpression was sufficient to cause an EAE-like disease, indicating a causatory role[201].

Additional support is provided by four different mouse models of MS, two EAE-like and two other models, that were shown to be alleviated by PADI inhibition (using 2CA, a pan-PADI inhibitor)[206,209,210]. A further structurally unrelated non-covalent inhibitor (with increased potency for PADI2, but with inhibitory activity on PADI1 and PADI4) also showed efficacy in the EAE model[211]. Most recently a fifth and new MS model, cuprizone autoimmune encephalitis (CAE), was developed where biochemical perturbation of MBP precedes an autoimmune inflammatory aspect of the disease, that drives a potent EAE-like disease in the absence of an exogenous antigen or antibodies[167]. Cuprizone, a copper chelator known to cause reversible demyelination (and used in remyelinating models of multiple sclerosis), was used briefly to perturb myelin structure, but without causing demyelination[167]. An immune stimulus was then used, but without the inclusion of exogenous myelin peptides (that are required in EAE models), to observe whether endogenous biochemically disrupted myelin might substitute for the exogenous peptides and drive a similar secondary inflammatory

demyelination[167]. The immune stimulus had no effect on untreated myelin, but drove a secondary inflammatory demyelination disease phenotype after cuprizone pre-treatment (CAE disease phenotype)[167]. Importantly, PADI inhibition (pan-PADI inhibition with BB-Cl-amidine and also with KP-302[211], a structurally unrelated inhibitor with 2-fold increased potency to PADI2) blocked the CAE phenotype, even when only administered during the initial cuprizone treatment[167]. These data point to a possible model of MS which posits that disrupted myelin and oligodendrocyte death drives a second inflammatory demyelinating part of disease[167,212]. It also shows the myelin perturbation that drove the subsequent inflammatory demyelination was dependent on PADI activity, confirming the earlier data showing the potential importance of PADI activity in disease aetiology[167].

Three lines of evidence are of particular importance: 1) increases in PADI expression and activity in normal appearing white matter from multiple sclerosis patients[213], 2) that multiple structurally unrelated PADI inhibitors are efficacious in models of EAE[206,209-211], especially in the most recent CAE model[167], and 3) the causative role for driving an EAE-type disease in PADI2 overexpressing mice[201], which show a concomitant increase in PADI4 expression. This suggests a direct causative role for PADI enzymes, and in particular for both PADI2 and PADI4, in multiple sclerosis. The hypothesis that citrullinated myelin precedes a secondary inflammatory demyelination is compelling, especially in light of the failure of direct immunomodulatory therapeutics, sometimes called the “inside out” view of MS. Even if remyelination is made easier, a secondary autoimmune attack will reverse this repair and regenerate symptoms.

Further work is required to determine precisely how citrullinated myelin components may be immunogenic and how PADIs impact on the human disease. PADI inhibitors, with a possible requirement for dual PADI2 and PADI4 inhibitors, appear to be highly promising candidates for therapeutic

intervention. Recent roles for PADI2 in oligodendrocytes will also need to be considered carefully with respect to therapeutic inhibition[165,208].

1.8.2 PADIs and Rheumatoid Arthritis

Citrullination also relates strongly to RA, in which immune cells attack cells that line the joints, with a long disease progression. PADI inhibition has been promising in disease models of RA and points potentially towards mechanism. Cl-amidine and PADI4 knockout reduced disease severity in the type II collagen induced and GPI induced arthritis models[202,214-216], but not in the K/BxN or collagen antibody-induced arthritis models[214,217]. The efficacy of PADI inhibition or knockout in the first two models is consistent with a role in autoimmune priming in which tolerance is broken as opposed to a role in the aspect of immune recruitment and joint destruction. Treatment with antibodies, rather than the antigen, bypasses the aspect of breaking of tolerance in terms of disease aetiology. Much of what is missing to resolve these observations lies in our incomplete understanding of how PADI4 protein becomes activated and how PADI4 plays a precise role in neutrophil biology.

Several connections between RA and citrullination have also been discovered. The first derives from genetic association studies between SNPs in PADI4 and RA disease; an initial study found SNPs in PADI4 to be significantly enriched in a Japanese RA cohort[218]. Although this could not be reproduced in studies using UK, French and Spanish RA cohorts[219-221], it was subsequently confirmed by multiple other cohorts[222-224] and in various meta-analyses including to the locus 1p36 which contains the PADIs[225-227]. The inconsistencies between studies were subsequently suggested to be due to insufficient power to detect the genetic association[225-227].

The second connection to citrullination has had widespread implications for diagnosis, if not yet fully for mechanism. Although little is understood about the cause of RA, one of the main diagnostic features is the presence of

specific autoantibodies in patient serum. Classically, rheumatoid factor was used to identify RA, but it is also present in many other diseases associated with chronic inflammation, and in up to 5-10% of healthy people in addition. Detecting the presence of anti-citrullinated protein antibodies (ACPAs) has emerged as a powerful and more specific diagnostic test for RA; RA is now typically classified between ACPA positive and negative forms in the clinic. Approximately 75% of patient cases present as ACPA positive[228,229]. Notably, citrullination appears not to be a bystander or secondary effect to the inflammation as ACPAs are highly specific for RA, appear very early (up to 9 years before the presentation of symptoms), and are associated with the most erosive cases of RA[228,229]. Gene environment interactions in RA also associate with citrullination such as smoking and the HLA genotype. The first ACPA epitope was found to be citrullinated fibrinogen, but not unmodified fibrinogen. Various anti-citrullinated protein epitopes have since been discovered.

The third connection derives from the role of PADI4 in neutrophils and NET formation. What is known conclusively is that neutrophils and activated neutrophils play an essential role in both the initiation and progression of RA[230]. The elucidation, firstly of the precise role for PADI4 in NET formation will be important to establish the connection strongly between citrullination and RA. Secondly it remains to be convincingly demonstrated that the role of PADI4 in NET formation is the same as that responsible for the connection of PADIs to RA[173,177]. The suggestion that cross-reactive antibodies appearing in the synovial fluid and found in RA increase the calcium sensitivity of PADI4[231] provides a mechanism in principle for how the two aspects could progressively exacerbate disease aetiology, but unambiguously establishing how these two aspects relate will be crucial moving forward.

The association of the related homocitrulline (carbamylated lysine) with RA provides a final interesting additional suggestion that the biochemistry of citrullinated proteins might be responsible for the mechanism that leads to

the breaking of tolerance[232-235]. Lysine carbamylation, which can be mediated non-enzymatically or by the byproduct of myeloperoxidase (MPO) is a post-translational modification that produces homocitrulline (chemically identical to citrulline, but with one additional carbon atom in the side chain)[232-235]. MPO converts thiocyanate to cyanate, which can enzymatically carbamylate lysine such as when MPO is released by activated neutrophils[232-235]. Alternatively, urea in the body is in equilibrium with cyanate, so if urea concentrations are high then non-enzymatic carbamylation of lysine residues can occur[232-235]. What is remarkable is that 8.75% of RA patients (comprising approximately a third of ACPA negative patients) present with autoantibodies to anti-carbamylated proteins[188,232-235]. That an enzymatically unrelated process gives rise to the same biochemical moiety with similar autoimmune consequences is suggestive that something about the biochemistry of citrullinated residues might be mechanistic for the process of breaking tolerance in RA. That citrullination has emerged as a hallmark for such a variety of autoimmune conditions would be indicative of this.

1.8.3 PADIs and cancer

PADIs have been found to play roles in cancer[127,128]. This has been primarily observed due to overexpression of PADIs in tumour contexts including breast cancer[236], prostate cancer}[237], colorectal cancer[238,239], leukaemia[240,241], and in late stage adenoma/carcinoma (such as esophageal squamous cell adenoma)[242,243]. Roles have been found firstly for their role as an epigenetic modifier in regulating gene expression and in creating a permissive decondensed chromatin environment mediated by histone citrullination. Secondly, roles have been found in promoting cancerous transformation both through effects of citrullination on oncogenes and tumour suppressors[244], but also in hormone dependent cancer activation (e.g. breast and prostate)[236,237]. Thirdly roles have been found for their regulation of inflammation and cell signalling in a cancer specific context. PADI inhibitors have shown initial promise in breast cancer and are toxic to various cancer cell lines. A particularly interesting example with therapeutic potential has connected the

roles of PADIs in autoimmunity to those in cancer contexts[123]. It was recognized that certain mutated antigens can be recognized by tumor reactive T cells[245,246] and that citrullinated antigens pertain to cancer contexts[247-249]. Presentation of citrullinated peptides on major histocompatibility complex class II (MHC class II) molecules in tumour cells could therefore provide a target for a CD4+ T cell mediated immunotherapeutic approach directed specifically against tumour cells[247-250]. This has been recently demonstrated for citrullinated enolase[250].

One of the most interesting connections to cancer possibly relates to the role of PADI4 in stem cells[9,241,251]. Analyzing normal haematopoietic stem cells (HSCs) and leukaemia progenitor cells revealed a signature of genes that were upregulated in both cell types and included PADI4[252]. Genes that regulate both self-renewal and oncogenesis are particularly interesting from the perspective of targeting cancer stem cells. This also would appear to be a particularly interesting area for future research in looking at any connections or mechanisms in common between the role of PADI4 in self-renewal or in induced pluripotent stem cells and the high expression of PADI4 that is seen in normal HSCs.

1.9 Targeting PADIs

Given the connections to disease and the tractability of PADIs as enzymes for small molecule perturbation, efforts have been made to target the PADIs[76,183,253,254]. The first PADI inhibitor to be described was paclitaxel, which independently from its role in microtubule stabilization, was found to weakly inhibit PADI2[255]. Other weak reversible PADI inhibitors have been discovered including streptomycin, chlortetracycline and minocycline[256]. The first major breakthrough in targeting PADIs was Cl-amidine, which was based on the success of irreversible inhibitors in targeting cysteine proteases[257-259]. Short synthetic peptides (benzoyl l-arginine ethylester and benzoyl l-arginine amide), which have been used in mechanistic studies of PADIs as efficient substrates, were modified by replacing the substrate arginine with a

haloacetamidine warhead[260,261]. The resultant molecule reacts covalently with the PADI active site cysteine[257-259]. This led to the generation of F-amidine, and then Cl-amidine, which was more effective[260,261], and have been widely used. More recently a modified more hydrophobic and cell permeable analogue of Cl-amidine, called BB-Cl-amidine, improved the half maximal inhibitory concentration (IC_{50}) in cells[262,263]. Not all PADIs are targeted as efficiently by Cl-amidine, but since such high concentrations (200 μM) are used to compensate for its poor cell permeability, it is generally considered to be a pan-PADI inhibitor. Some progress has also been made in generating reversible inhibitors. The first showed selectivity for PADI4, targeting a loop in the calcium-unbound apo-structure[264]. A second reversible inhibitor with some selectivity for PADI2 (but which still inhibits PADI1 and PADI4) has also been described[211]. This will be addressed further in the short introduction to Chapter 5.

1.10 References for Chapter 1

1. Waddington, C. H. Towards a Theoretical Biology. *Nature* **218,** 525–527 (1968).
2. Wu, C. T. & Morris, J. R. Genes, Genetics, and Epigenetics: A Correspondence. *Science* **293,** 1103–1105 (2001).
3. Dupont, C., Armant, D. R. & Brenner, C. A. Epigenetics: definition, mechanisms and clinical perspective. *Semin. Reprod. Med.* **27,** 351–357 (2009).
4. Gurdon, J. B., Elsdale, T. R. & Fischberg, M. Sexually Mature Individuals of Xenopus laevis from the Transplantation of Single Somatic Nuclei. *, Published online: 05 July 1958; | doi:10.1038/182064a0* **182,** 64–65 (1958).
5. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126,** 663–676 (2006).
6. Tanabe, K. *et al.* Transdifferentiation of human adult peripheral blood T cells into neurons. *PNAS* **115,** 6470–6475 (2018).
7. Wang, Y. *et al.* Human PAD4 regulates histone arginine methylation levels via demethylimination. *Science* **306,** 279–283 (2004).
8. Wang, Y. *et al.* Histone hypercitrullination mediates chromatin decondensation and neutrophil extracellular trap formation. *J Cell Biol* **184,** 205–213 (2009).
9. Christophorou, M. A. *et al.* Citrullination regulates pluripotency and histone H1 binding to chromatin. *Nature* **507,** 104–108 (2014).
10. Consortium, I. H. G. S. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
11. Pertea, M. *et al.* CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biol.* **19,** 208 (2018).
12. Willyard, C. New human gene tally reignites debate. *Nature* **558,** 354–356 (2018).
13. Walsh, C. T., Garneau-Tsodikova, S. & Gatto, G. J. Protein posttranslational modifications: the chemistry of proteome diversifications. *Angew. Chem. Int. Ed. Engl.* **44,** 7342–7372 (2005).
14. Kennani, El, S., Crespo, M., Govin, J. & Pflieger, D. Proteomic Analysis of Histone Variants and Their PTMs: Strategies and Pitfalls. *Proteomes* **6,** 29 (2018).
15. Strahl, B. D. & Allis, C. D. The language of covalent histone modifications. *Nature* **403,** 41–45 (2000).
16. Jenuwein, T. & Allis, C. D. Translating the histone code. *Science* **293,** 1074–1080 (2001).
17. Ciehanover, A., Hod, Y. & Hershko, A. A heat-stable polypeptide component of an ATP-dependent proteolytic system from reticulocytes. *Biochemical and Biophysical Research Communications* **81,** 1100–1105 (1978).
18. Varshavsky, A. The N-end rule pathway and regulation by proteolysis. *Protein Sci.* **20,** 1298–1345 (2011).
19. Duan, G. & Walther, D. The roles of post-translational modifications in the context of protein interaction networks. *PLOS Computational Biology* **11,** e1004049 (2015).
20. Harauz, G. *et al.* Myelin basic protein-diverse conformational states of an intrinsically unstructured protein and its roles in myelin assembly and multiple sclerosis. *Micron* **35,** 503–542 (2004).
21. Zhang, C. *et al.* Myelin Basic Protein Undergoes a Broader Range of Modifications in Mammals than in Lower Vertebrates. *J. Proteome Res.* **11,** 4791–4802 (2012).
22. Cohen, P. The regulation of protein function by multisite phosphorylation – a 25 year update. *Trends in Biochemical Sciences* **25,** 596–601 (2000).
23. Sabatini, D. M. Twenty-five years of mTOR: Uncovering the link from nutrients to growth. *PNAS* **114,** 11818–11825 (2017).
24. Schofield, C. J. & Ratcliffe, P. J. Oxygen sensing by HIF hydroxylases. *Nat. Rev. Mol. Cell Biol.* **5,** 343–354 (2004).
25. Farah, C., Michel, L. Y. M. & Balligand, J.-L. Nitric oxide signalling in cardiovascular health and disease. *Nat Rev Cardiol* **15,** 292–316 (2018).
26. Mitsui, A. *et al.* Strategy by which nitrogen-fixing unicellular cyanobacteria grow photoautotrophically. *Nature* **323,** 720–722 (1986).

27. Cohen, S. E. & Golden, S. S. Circadian Rhythms in Cyanobacteria. *Microbiol. Mol. Biol. Rev.* **79,** 373–385 (2015).
28. Hastings, M. H., Maywood, E. S. & Brancaccio, M. Generation of circadian rhythms in the suprachiasmatic nucleus. *Nat Rev Neurosci* **19,** 453–469 (2018).
29. Piccolo, S., Dupont, S. & Cordenonsi, M. The Biology of YAP/TAZ: Hippo Signaling and Beyond. *Physiol. Rev.* **94,** 1287–1312 (2014).
30. Nusse, R. & Clevers, H. Wnt/β-Catenin Signaling, Disease, and Emerging Therapeutic Modalities. *Cell* **169,** 985–999 (2017).
31. Cohen, P. The role of protein phosphorylation in human health and disease. *The FEBS Journal* **268,** 5001–5010 (2001).
32. Hershko, A. & Ciechanover, A. The Ubiquitin System. *Annu. Rev. Biochem.* **67,** 425–479 (1998).
33. Swatek, K. N. & Komander, D. Ubiquitin modifications. *Cell Res.* **26,** 399–422 (2016).
34. Hunter, T. The age of crosstalk: Phosphorylation, ubiquitination, and beyond. *Molecular Cell* **28,** 730–738 (2007).
35. Clevers, H. & Nusse, R. Wnt/β-Catenin Signaling and Disease. *Cell* **149,** 1192–1205 (2012).
36. Stamos, J. L. & Weis, W. I. The beta-Catenin Destruction Complex. *Cold Spring Harb Perspect Biol* **5,** –a007898 (2013).
37. Fuhrmann, J., Clancy, K. W. & Thompson, P. R. Chemical Biology of Protein Arginine Modifications in Epigenetic Regulation. *Chem. Rev.* **115,** 5413–5461 (2015).
38. Koga, Y. Study report on the constituents of squeezed watermelon. *Tokyo Kagaku Kaishi [Journal of the Tokyo Chemical Society]* **35,** 519–528 (1914).
39. Schriek, S., Rueckert, C., Staiger, D., Pistorius, E. K. & Michel, K.-P. Bioinformatic evaluation of L-arginine catabolic pathways in 24 cyanobacteria and transcriptional analysis of genes encoding enzymes of L-arginine catabolism in the cyanobacterium Synechocystis sp PCC 6803. *BMC Genomics* **8,** (2007).
40. Jackson, M. J. & Beaudet, A. L. Mammalian urea cycle enzymes. *Annual review of genetics* (1986).
41. Morris, S. M. Regulation of Enzymes of the Urea Cycle and Arginine Metabolism. *Annu. Rev. Nutr.* **22,** 87–105 (2002).
42. Griffith, O. W. & Stuehr, D. J. Nitric oxide synthases: properties and catalytic mechanism. *Annual Review of Physiology* (1995).
43. Stuehr, D. J. Mammalian nitric oxide synthases. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **1411,** 217–230 (1999).
44. Lundberg, J. O., Weitzberg, E. & Gladwin, M. T. The nitrate–nitrite–nitric oxide pathway in physiology and therapeutics. *Nature Reviews Drug Discovery 2004 3:6* **7,** 156–167 (2008).
45. Leiper, J. M. *et al.* Identification of two human dimethylarginine dimethylaminohydrolases with distinct tissue distributions and homology with microbial arginine deiminases. *Biochem. J.* **343 Pt 1,** 209–214 (1999).
46. Murray-Rust, J. *et al.* Structural insights into the hydrolysis of cellular nitric oxide synthase inhibitors by dimethylarginine dimethylaminohydrolase. *Nat. Struct. Biol.* **8,** 679–683 (2001).
47. Wada, M. Uber Citrullin, eine neue Aminosaure im PreBsaft der Wassermelone, Citrullus vulgaris schrad. *Biochemische Zeitschrift* **224,** 420–429 (1930).
48. Wada, M. Isolierung des Citrullins (δ-Carbamido-ornithin) aus tryptischen Verdauungsprodukten des Caseins. *Biochemische Zeitschrift* **257,** 1–7 (1933).
49. Darrah, E. & Andrade, F. Rheumatoid arthritis and citrullination. *Current Opinion in Rheumatology* **30,** 72–78 (2018).
50. Rogers, G. E. & Simmonds, D. H. Content of Citrulline and Other Amino-Acids in a Protein of Hair Follicles. *Nature* **182,** 186–187 (1958).
51. Rogers, G. E. Occurrence of Citrulline in Proteins. *Nature* **194,** 1149–1151 (1962).
52. Finch, P. R., Wood, D. D. & Moscarello, M. A. The presence of citrulline in a myelin protein fraction. *FEBS Letters* **15,** 145–148 (1971).
53. Rogers, G. E., Harding, H. W. J. & Llewellyn-Smith, I. J. The origin of citrulline-

containing proteins in the hair follicle and the chemical nature of trichohyalin, an intracellular precursor. *Biochimica et Biophysica Acta (BBA) - Protein Structure* **495,** 159–175 (1977).
54. Fujisaki, M. & Sugawara, K. Properties of Peptidylarginine Deiminase From the Epidermis of Newborn Rats. *J Biochem* **89,** 257–263 (1981).
55. Nicholas, A. P. & Bhattacharya, S. K. *Protein deimination in human health and disease*. (Springer New York, 2014).
56. Balandraud, N. *et al.* A rigorous method for multigenic families' functional annotation: the peptidyl arginine deiminase (PADs) proteins family example. *BMC Genomics* **6,** 153 (2005).
57. Wang, S. & Wang, Y. Peptidylarginine deiminases in citrullination, gene regulation, health and pathogenesis. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1829,** 1126–1135 (2013).
58. McGraw, W. T., Potempa, J., Farley, D. & Travis, J. Purification, Characterization, and Sequence Analysis of a Potential Virulence Factor from Porphyromonas gingivalis, Peptidylarginine Deiminase. *Infect. Immun.* **67,** 3248–3256 (1999).
59. Carolina Touz, M. *et al.* Arginine deiminase has multiple regulatory roles in the biology of Giardia lamblia. *J Cell Sci* **121,** 2930–2938 (2008).
60. Goulas, T. *et al.* Structure and mechanism of a bacterial host-protein citrullinating virulence factor, Porphyromonas gingivalis peptidylarginine deiminase. *Sci Rep* **5,** 11969 (2015).
61. Shirai, H., Blundell, T. L. & Mizuguchi, K. A novel superfamily of enzymes that catalyze the modification of guanidino groups. *Trends in Biochemical Sciences* **26,** 465–468 (2001).
62. Montgomery, A. B. *et al.* Crystal structure of Porphyromonas gingivalis peptidylarginine deiminase: implications for autoimmunity in rheumatoid arthritis. *Ann Rheum Dis* **75,** 1255–1261 (2016).
63. Stobernack, T. *et al.* The Extracellular Proteome and Citrullinome of the Oral Pathogen Porphyromonas gingivalis. *J. Proteome Res.* **15,** acs.jproteome.6b00634–4543 (2016).
64. Li, Z. *et al.* Mechanisms of catalysis and inhibition operative in the arginine deiminase from the human pathogen Giardia lamblia. *Bioorganic Chemistry* **37,** 149–161 (2009).
65. Linsky, T. & Fast, W. Mechanistic similarity and diversity among the guanidine-modifying members of the pentein superfamily. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **1804,** 1943–1953 (2010).
66. Groft, C. M., Beckmann, R., Sali, A. & Burley, S. K. Crystal structures of ribosome anti-association factor IF6. *Nature Structural & Molecular Biology* **7,** 1156–1164 (2000).
67. Paoli, M. An elusive propeller-like fold. *Nature Structural & Molecular Biology* **8,** 744–744 (2001).
68. Shirai, H., Mokrab, Y. & Mizuguchi, K. The guanidino group modifying enzymes: Structural basis for their diversity and commonality. *Proteins: Structure, Function, and Bioinformatics* **64,** 1010–1023 (2006).
69. György, B., Tóth, E., Tarcsa, E., Falus, A. & Buzás, E. I. Citrullination: A posttranslational modification in health and disease. *The International Journal of Biochemistry & Cell Biology* **38,** 1662–1677 (2006).
70. Arita, K. *et al.* Structural basis for Ca2+-induced activation of human PAD4. *Nature Structural & Molecular Biology* **11,** 777–783 (2004).
71. Slade, D. J. *et al.* Protein arginine deiminase 2 binds calcium in an ordered fashion: implications for inhibitor design. *ACS Chem. Biol.* **10,** 1043–1053 (2015).
72. Stone, E. M., Costello, A. L., Tierney, D. L. & Fast, W. Substrate-Assisted Cysteine Deprotonation in the Mechanism of Dimethylargininase (DDAH) from Pseudomonas aeruginosa†. *Biochemistry* **45,** 5618–5630 (2006).
73. Knuckley, B., Bhatia, M. & Thompson, P. R. Protein arginine deiminase 4: evidence for a reverse protonation mechanism. *Biochemistry* **46,** 6578–6587 (2007).
74. Knuckley, B. *et al.* Substrate specificity and kinetic studies of PADs 1, 3, and 4 identify potent and selective inhibitors of protein arginine deiminase 3. *Biochemistry*

49, 4852–4863 (2010).
75. Dreyton, C. J., Knuckley, B., Jones, J. E., Lewallen, D. M. & Thompson, P. R. Mechanistic studies of protein arginine deiminase 2: evidence for a substrate-assisted mechanism. *Biochemistry* **53,** 4426–4433 (2014).
76. Mondal, S. & Thompson, P. R. Protein Arginine Deiminases (PADs): Biochemistry and Chemical Biology of Protein Citrullination. *Acc. Chem. Res.* **52,** 818–832 (2019).
77. Calnan, B. J., Tidor, B., Biancalana, S., Hudson, D. & Frankel, A. D. Arginine-mediated RNA recognition: the arginine fork. *Science* **252,** 1167–1171 (1991).
78. Jones, S. & Thornton, J. M. Principles of protein-protein interactions. *PNAS* **93,** 13–20 (1996).
79. Jones, S. & Thornton, J. M. Analysis of protein-protein interaction sites using surface patches. *Journal of Molecular Biology* **272,** 121–132 (1997).
80. Luscombe, N. M., Laskowski, R. A. & Thornton, J. M. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucl. Acids Res.* **29,** 2860–2874 (2001).
81. Truant, R. & Cullen, B. R. The Arginine-Rich Domains Present in Human Immunodeficiency Virus Type 1 Tat and Rev Function as Direct Importin β-Dependent Nuclear Localization Signals. *Mol. Cell. Biol.* **19,** 1210–1217 (1999).
82. Khan, M. S. *et al.* Serpin Inhibition Mechanism: A Delicate Balance between Native Metastable State and Polymerization. *J Amino Acids* **2011,** 606797–10 (2011).
83. Qamar, S. *et al.* FUS Phase Separation Is Modulated by a Molecular Chaperone and Methylation of Arginine Cation-π Interactions. *Cell* **173,** 720–734.e15 (2018).
84. Snijders, A. P. *et al.* Arginine methylation and citrullination of splicing factor proline- and glutamine-rich (SFPQ/PSF) regulates its association with mRNA. *RNA* **21,** 347–359 (2015).
85. Guo, Q. & Fast, W. Citrullination of Inhibitor of Growth 4 (ING4) by Peptidylarginine Deminase 4 (PAD4) Disrupts the Interaction between ING4 and p53. *J. Biol. Chem.* **286,** 17069–17078 (2011).
86. Tanikawa, C. *et al.* Regulation of protein Citrullination through p53/PADI4 network in DNA damage response. *Cancer Res.* **69,** 8761–8769 (2009).
87. Stadler, S. C. *et al.* Dysregulation of PAD4-mediated citrullination of nuclear GSK3β activates TGF-β signaling and induces epithelial-to-mesenchymal transition in breast cancer cells. *PNAS* **110,** 11851–11856 (2013).
88. Sun, B. *et al.* Citrullination of NF-kB p65 enhances its nuclear localization and TLR-induced expression of IL-1β and TNFα. *Science immunology* **2,** eaal3062 (2017).
89. Moscarello, M. A., Wood, D. D., Ackerley, C. & Boulias, C. Myelin in multiple sclerosis is developmentally immature. *Journal of Clinical Investigation* **94,** 146–154 (1994).
90. Pritzker, L. B., Joshi, S., Jessica J Gowan, Harauz, G. & Moscarello, M. A. Deimination of Myelin Basic Protein. 1. Effect of Deimination of Arginyl Residues of Myelin Basic Protein on Its Structure and Susceptibility to Digestion by Cathepsin D†. *Biochemistry* **39,** 5374–5381 (2000).
91. Tanikawa, C. *et al.* Citrullination of RGG Motifs in FET Proteins by PAD4 Regulates Protein Aggregation and ALS Susceptibility. *Cell Rep* **22,** 1473–1483 (2018).
92. Trentini, D. B. *et al.* Arginine phosphorylation marks proteins for degradation by a Clp protease. *Nature* **539,** 48–53 (2016).
93. Hardman, G. *et al.* Strong anion exchange mediated phosphoproteomics reveals extensive human non canonical phosphorylation. *The EMBO Journal* **13,** Unit 13 15–24 (2019).
94. Lüscher, B. *et al.* ADP-Ribosylation, a Multifaceted Posttranslational Modification Involved in the Control of Cell Physiology in Health and Disease. *Chem. Rev.* **118,** 1092–1136 (2017).
95. Galligan, J. J. *et al.* Methylglyoxal-derived posttranslational arginine modifications are abundant histone marks. *PNAS* **115,** 9228–9233 (2018).
96. Wilkins, S. E. *et al.* JMJD5 is a human arginyl C-3 hydroxylase. *Nat Commun* **9,** 1–12 (2018).

97. Larsen, S. C. *et al.* Proteome-wide analysis of arginine monomethylation reveals widespread occurrence in human cells. *Sci Signal* **9,** –rs9 (2016).
98. Blanc, R. S. & Richard, S. Arginine Methylation: The Coming of Age. *Molecular Cell* **65,** 8–24 (2017).
99. Chen, D. *et al.* Regulation of Transcription by a Protein Methyltransferase. *Science* **284,** 2174–2177 (1999).
100. Schurter, B. T. *et al.* Methylation of histone H3 by coactivator-associated arginine methyltransferase 1. *Biochemistry* **40,** 5747–5756 (2001).
101. Ma, H. *et al.* Hormone-dependent, CARM1-directed, arginine-specific methylation of histone H3 on a steroid-regulated promoter. *Current Biology* **11,** 1981–1985 (2001).
102. Bauer, U. M., Daujat, S., Nielsen, S. J., Nightingale, K. & Kouzarides, T. Methylation at arginine 17 of histone H3 is linked to gene activation. *EMBO reports* **3,** 39–44 (2002).
103. Thompson, P. R. & Fast, W. Histone citrullination by protein arginine deiminase: Is arginine methylation a green light or a roadblock? *ACS Chem. Biol.* **1,** 433–441 (2006).
104. Wang, H. *et al.* Methylation of Histone H4 at Arginine 3 Facilitating Transcriptional Activation by Nuclear Hormone Receptor. *Science* **293,** 853–857 (2001).
105. Cuthbert, G. L. *et al.* Histone Deimination Antagonizes Arginine Methylation. *Cell* **118,** 545–553 (2004).
106. Sharma, P. *et al.* Arginine Citrullination at the C-Terminal Domain Controls RNA Polymerase II Transcription. *Molecular Cell* (2018). doi:10.1016/j.molcel.2018.10.016
107. Guo, Q., Bedford, M. T. & Fast, W. Discovery of peptidylarginine deiminase-4 substrates by protein array: antagonistic citrullination and methylation of human ribosomal protein S2. *Mol Biosyst* **7,** 2286–2295 (2011).
108. Hidaka, Y., Hagiwara, T. & Yamada, M. Methylation of the guanidino group of arginine residues prevents citrullination by peptidylarginine deiminase IV. *FEBS Letters* **579,** 4088–4092 (2005).
109. Böttger, A., Islam, M. S., Chowdhury, R., Schofield, C. J. & Wolf, A. The oxygenase Jmjd6--a case study in conflicting assignments. *Biochem. J.* **468,** 191–202 (2015).
110. Walport, L. J. *et al.* Arginine demethylation is catalysed by a subset of JmjC histone lysine demethylases. *Nat Commun* **7,** 1–12 (2016).
111. Chory, E. J. *et al.* Nucleosome Turnover Regulates Histone Methylation Patterns over the Genome. *Molecular Cell* **73,** 61–72.e3 (2019).
112. Shi, Y. *et al.* Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. *Cell* **119,** 941–953 (2004).
113. Tsukada, Y.-I. *et al.* Histone demethylation by a family of JmjC domain-containing proteins. *Nature* **439,** 811–816 (2005).
114. Shi, Y. & Whetstine, J. R. Dynamic Regulation of Histone Lysine Methylation by Demethylases. *Molecular Cell* **25,** 1–14 (2007).
115. Beurel, E., Grieco, S. F. & Jope, R. S. Glycogen synthase kinase-3 (GSK3): Regulation, actions, and diseases. *Pharmacology & Therapeutics* **148,** 114–131 (2015).
116. Bautista, S. J. *et al.* mTOR complex 1 controls the nuclear localization and function of glycogen synthase kinase 3β. *J. Biol. Chem.* **293,** 14723–14739 (2018).
117. Zhang, X. *et al.* Nuclear localization signal of ING4 plays a key role in its binding to p53. *Biochemical and Biophysical Research Communications* **331,** 1032–1038 (2005).
118. Koziel, J. *et al.* Citrullination Alters Immunomodulatory Function of LL-37 Essential for Prevention of Endotoxin-Induced Sepsis. *J Immunol* 1303062 (2014). doi:10.4049/jimmunol.1303062
119. Wong, A. *et al.* A Novel Biological Role for Peptidyl-Arginine Deiminases: Citrullination of Cathelicidin LL-37 Controls the Immunostimulatory Potential of Cell-Free DNA. *J Immunol* **200,** ji1701391–2340 (2018).
120. Loos, T. *et al.* Citrullination of CXCL10 and CXCL11 by peptidylarginine deiminase: a naturally occurring posttranslational modification of chemokines and new

dimension of immunoregulation. *Blood* **112,** 2648–2656 (2008).
121. Proost, P. *et al.* Citrullination of CXCL8 by peptidylarginine deiminase alters receptor usage, prevents proteolysis, and dampens tissue inflammation. *J Exp Med* **205,** 2085–2097 (2008).
122. Ordonez, A. *et al.* Effect of citrullination on the function and conformation of antithrombin. *FEBS Journal* **276,** 6763–6772 (2009).
123. Ordóñez, A. *et al.* Increased levels of citrullinated antithrombin in plasma of patients with rheumatoid arthritis and colorectal adenocarcinoma determined by a newly developed ELISA using a specific monoclonal antibody. *Thromb Haemost* **104,** 1143–1149 (2010).
124. Tilvawala, R. *et al.* The Rheumatoid Arthritis-Associated Citrullinome. *Cell Chem Biol* **25,** 691–+ (2018).
125. Law, R. H. *et al.* An overview of the serpin superfamily. *Genome Biol.* **7,** 1–11 (2006).
126. Sanrattana, W., Maas, C. & de Maat, S. SERPINs-From Trap to Treatment. *Front Med (Lausanne)* **6,** 25 (2019).
127. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144,** 646–674 (2011).
128. Yuzhalin, A. E. Citrullination in Cancer. *Cancer Res.* **79,** 1274–1284 (2019).
129. Brangwynne, C. P. *et al.* Germline P Granules Are Liquid Droplets That Localize by Controlled Dissolution/Condensation. *Science* **324,** 1729–1732 (2009).
130. Banani, S. F., Lee, H. O., Hyman, A. A. & Rosen, M. K. Biomolecular condensates: organizers of cellular biochemistry. *Nat. Rev. Mol. Cell Biol.* **18,** 285–298 (2017).
131. Hyman, A. A., Weber, C. A. & Jülicher, F. Liquid-Liquid Phase Separation in Biology. *Annu. Rev. Cell Dev. Biol.* **30,** 39–58 (2014).
132. Thandapani, P., O'Connor, T. R., Bailey, T. L. & Richard, S. Defining the RGG/RG motif. *Molecular Cell* **50,** 613–623 (2013).
133. Nozawa, R.-S. *et al.* SAF-A Regulates Interphase Chromosome Structure through Oligomerization with Chromatin-Associated RNAs. *Cell* **169,** 1214–1227.e18 (2017).
134. Strom, A. R. *et al.* Phase separation drives heterochromatin domain formation. *Nature* **547,** 241–+ (2017).
135. Larson, A. G. *et al.* Liquid droplet formation by HP1α suggests a role for phase separation in heterochromatin. *Nature* **547,** 236–240 (2017).
136. Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K. & Sharp, P. A. A Phase Separation Model for Transcriptional Control. *Cell* **169,** 13–23 (2017).
137. Harlen, K. M. & Churchman, L. S. The code and beyond: transcription regulation by the RNA polymerase II carboxy-terminal domain. *Nat. Rev. Mol. Cell Biol.* **18,** 263–273 (2017).
138. Burke, K. A., Janke, A. M., Rhine, C. L. & Fawzi, N. L. Residue-by-Residue View of In Vitro FUS Granules that Bind the C-Terminal Domain of RNA Polymerase II. *Molecular Cell* **60,** 231–241 (2015).
139. Kwon, I. *et al.* Phosphorylation-Regulated Binding of RNA Polymerase II to Fibrous Polymers of Low-Complexity Domains. *Cell* **155,** 1049–1060 (2013).
140. Janke, A. M. *et al.* Lysines in the RNA Polymerase II C-Terminal Domain Contribute to TAF15 Fibril Recruitment. *Biochemistry* **57,** 2549–2563 (2018).
141. Boehning, M. *et al.* RNA polymerase II clustering through carboxy-terminal domain phase separation. *Nature Structural & Molecular Biology* **25,** 833–840 (2018).
142. Corden, J. L. An Arginine Nexus in the RNA Polymerase II CTD. *Molecular Cell* **73,** 3–4 (2019).
143. Chavanas, S. *et al.* Comparative analysis of the mouse and human peptidylarginine deiminase gene clusters reveals highly conserved non-coding segments and a new human gene, PADI6. *Gene* **330,** 19–27 (2004).
144. Darrah, E., Rosen, A., Giles, J. T. & Andrade, F. Peptidylarginine deiminase 2, 3 and 4 have distinct specificities against cellular substrates: novel insights into autoantigen selection in rheumatoid arthritis. *Ann Rheum Dis* **71,** 92–98 (2012).
145. Assohou-Luty, C. *et al.* The human peptidylarginine deiminases type 2 and type 4 have distinct substrate specificities. *Biochim. Biophys. Acta* **1844,** 829–836 (2014).

146. Olson, J. S., Lubner, J. M., Meyer, D. J. & Grant, J. E. An in silico analysis of primary and secondary structure specificity determinants for human peptidylarginine deiminase types 2 and 4. *Computational Biology and Chemistry* **70,** 107–115 (2017).
147. Tarcsa, E. *et al.* Protein Unfolding by Peptidylarginine Deiminase. *J. Biol. Chem.* **271,** 30709–30716 (1996).
148. Tarcsa, E. *et al.* The Fate of Trichohyalin: Sequential Post-Translational Modifications by Peptidyl-Arginine Deiminase and Transglutaminases. *J. Biol. Chem.* **272,** 27893–27901 (1997).
149. Steinert, P. M., Parry, D. A. D. & Marekov, L. N. Trichohyalin Mechanically Strengthens the Hair Follicle. *J. Biol. Chem.* **278,** 41409–41419 (2003).
150. Traub, P. & Vorgias, C. E. Differential Effect of Arginine Modification with 1,2-Cyclohexanedione on the Capacity of Vimentin and Desmin to Assemble Into Intermediate Filaments and to Bind to Nucleic-Acids. *J Cell Sci* **65,** 1–20 (1984).
151. Inagaki, M., Nishi, Y., Nishizawa, K., Matsuyama, M. & Sato, C. Site-specific phosphorylation induces disassembly of vimentin filaments in vitro. *Nature* **328,** 649–652 (1987).
152. Inagaki, M., Takahara, H., Nishi, Y., Sugawara, K. & Sato, C. Ca2+-dependent deimination-induced disassembly of intermediate filaments involves specific modification of the amino-terminal head domain. *J. Biol. Chem.* **264,** 18119–18127 (1989).
153. Asaga, H., Yamada, M. & Senshu, T. Selective Deimination of Vimentin in Calcium Ionophore-Induced Apoptosis of Mouse Peritoneal Macrophages. *Biochemical and Biophysical Research Communications* **243,** 641–646 (1998).
154. Basmanav, F. B. U. *et al.* Mutations in Three Genes Encoding Proteins Involved in Hair Shaft Formation Cause Uncombable Hair Syndrome. *The American Journal of Human Genetics* **99,** 1292–1304 (2016).
155. Liu, F. *et al.* Meta-analysis of genome-wide association studies identifies 8 novel loci involved in shape variation of human head hair. *Human Molecular Genetics* **27,** 559–575 (2018).
156. Morgan, M. D. *et al.* Genome-wide study of hair colour in UK Biobank explains most of the SNP heritability. *Nat Commun* **9,** (2018).
157. Kan, R. *et al.* Potential role for PADI-mediated histone citrullination in preimplantation development. *BMC Dev. Biol.* **12,** (2012).
158. Esposito, G. *et al.* Peptidylarginine deiminase (PAD) 6 is essential for oocyte cytoskeletal sheet formation and female fertility. *Mol. Cell. Endocrinol.* **273,** 25–31 (2007).
159. Yurttas, P. *et al.* Role for PADI6 and the cytoplasmic lattices in ribosomal storage in oocytes and translational control in the early mouse embryo. *Development* **135,** 2627–2636 (2008).
160. Xu, Y. *et al.* Mutations in PADI6 Cause Female Infertility Characterized by Early Embryonic Arrest. *The American Journal of Human Genetics* **99,** 744–752 (2016).
161. Maddirevula, S. *et al.* The human knockout phenotype of PADI6 is female sterility caused by cleavage failure of their fertilized eggs. *Clin. Genet.* **91,** 344–345 (2017).
162. Qian, J. *et al.* Biallelic PADI6 variants linking infertility, miscarriages, and hydatidiform moles. *Eur. J. Hum. Genet.* **26,** 1007–1013 (2018).
163. Snow, A. J. *et al.* Phosphorylation-dependent interaction of tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein (YWHA) with PADI6 following oocyte maturation in mice. *Biol Reprod* **79,** 337–347 (2008).
164. Rose, R., Rose, M. & Ottmann, C. Identification and structural characterization of two 14-3-3 binding sites in the human peptidylarginine deiminase type VI. *Journal of Structural Biology* **180,** 65–72 (2012).
165. Falcao, A. M. *et al.* Disease-specific oligodendrocyte lineage cells arise in multiple sclerosis. *Nat Med* **24,** 1837–+ (2018).
166. Wood, D. D., Moscarello, M. A., Bilbao, J. M. & O'Connors, P. Acute multiple sclerosis (marburg type) is associated with developmentally immature myelin basic protein. *Annals of Neurology* **40,** 18–24 (1996).
167. Caprariello, A. V. *et al.* Biochemically altered myelin triggers autoimmune

demyelination. *PNAS* **115,** 5528–5533 (2018).
168. Harauz, G. & Musse, A. A. A Tale of Two Citrullines—Structural and Functional Aspects of Myelin Basic Protein Deimination in Health and Disease. *Neurochem Res* **32,** 137–158 (2007).
169. Tauber, A. I. Metchnikoff and the phagocytosis theory. *Nat. Rev. Mol. Cell Biol.* **4,** 897–901 (2003).
170. Brinkmann, V. *et al.* Neutrophil extracellular traps kill bacteria. *Science* **303,** 1532–1535 (2004).
171. Li, P. *et al.* PAD4 is essential for antibacterial innate immunity mediated by neutrophil extracellular traps. *J Exp Med* **207,** 1853–1862 (2010).
172. Hemmers, S., Teijaro, J. R., Arandjelovic, S. & Mowen, K. A. PAD4-mediated neutrophil extracellular trap formation is not required for immunity against influenza infection. *PLoS ONE* **6,** e22043 (2011).
173. Liu, Y. *et al.* Peptidylarginine deiminases 2 and 4 modulate innate and adaptive immune responses in TLR-7 dependent lupus. *JCI Insight* **3,** (2018).
174. Kenny, E. F. *et al.* Diverse stimuli engage different neutrophil extracellular trap pathways. *eLife* **6,** 178 (2017).
175. de Bont, C. M., Koopman, W. J. H., Boelens, W. C. & Pruijn, G. J. M. Stimulus-dependent chromatin dynamics, citrullination, calcium signalling and ROS production during NET formation. *Biochim. Biophys. Acta* **1865,** 1621–1629 (2018).
176. Neeli, I., Khan, S. N. & Radic, M. Histone deimination as a response to inflammatory stimuli in neutrophils. *J Immunol* **180,** 1895–1902 (2008).
177. Bawadekar, M. *et al.* Peptidylarginine deiminase 2 is required for tumor necrosis factor alpha-induced citrullination and arthritis, but not neutrophil extracellular trap formation. *Journal of Autoimmunity* **80,** 39–47 (2017).
178. Theunissen, T. W. *et al.* Nanog Overcomes Reprogramming Barriers and Induces Pluripotency in Minimal Conditions. *Current Biology* **21,** 65–71 (2011).
179. Zhang, X. *et al.* Peptidylarginine deiminase 1-catalyzed histone citrullination is essential for early embryo development. *Sci Rep* **6,** 38727 (2016).
180. Leshner, M. *et al.* PAD4 mediated histone hypercitrullination induces heterochromatin decondensation and chromatin unfolding to form neutrophil extracellular trap-like structures. *Front Immunol* **3,** 307 (2012).
181. Fuchs, T. A. *et al.* Novel cell death program leads to neutrophil extracellular traps. *J Cell Biol* **176,** 231–241 (2007).
182. Chen, X. *et al.* ATAC-see reveals the accessible genome by transposase-mediated imaging and sequencing. *Nature Methods* (2016). doi:10.1038/nmeth.4031
183. Lewis, H. D. & Nacht, M. iPAD or PADi—'tablets' with therapeutic disease potential? *Current Opinion in Chemical Biology* **33,** 169–178 (2016).
184. Ishida-Yamamoto, A. *et al.* Decreased deiminated keratin K1 in psoriatic hyperproliferative epidermis. *Journal of Investigative Dermatology* **114,** 701–705 (2000).
185. Lin, A. M. *et al.* Mast cells and neutrophils release IL-17 through extracellular trap formation in psoriasis. *J. Immunol.* **187,** 490–500 (2011).
186. Villanueva, E. *et al.* Netting neutrophils induce endothelial damage, infiltrate tissues, and expose immunostimulatory molecules in systemic lupus erythematosus. *J. Immunol.* **187,** 538–552 (2011).
187. Carmona-Rivera, C., Zhao, W., Yalavarthi, S. & Kaplan, M. J. Neutrophil extracellular traps induce endothelial dysfunction in systemic lupus erythematosus through the activation of matrix metalloproteinase-2. *Ann Rheum Dis* **74,** 1417–1424 (2015).
188. Ziegelasch, M. *et al.* Antibodies against carbamylated proteins and cyclic citrullinated peptides in systemic lupus erythematosus: results from two well-defined European cohorts. *Arthritis Res. Ther.* **18,** 289 (2016).
189. McGinty, J. W. *et al.* Recognition of Posttranslationally Modified GAD65 Epitopes in Subjects With Type 1 Diabetes. *Diabetes* **63,** 3033–3040 (2014).
190. Wong, S. L. *et al.* Diabetes primes neutrophils to undergo NETosis, which impairs wound healing. *Nat Med* **21,** 815–819 (2015).
191. Buitinga, M. *et al.* Inflammation-Induced Citrullinated Glucose-Regulated Protein 78

Elicits Immune Responses in Human Type 1 Diabetes. *Diabetes* **67,** 2337–2348 (2018).
192. Bhattacharya, S. K. *et al.* Proteomics Implicates Peptidyl Arginine Deiminase 2 and Optic Nerve Citrullination in Glaucoma Pathogenesis. *Invest. Ophthalmol. Vis. Sci.* **47,** 2508–2514 (2006).
193. Martinod, K. *et al.* Neutrophil histone modification by peptidylarginine deiminase 4 is critical for deep vein thrombosis in mice. *PNAS* **110,** 8674–8679 (2013).
194. Li, Y. *et al.* Citrullinated histone H3: a novel target for the treatment of sepsis. *Surgery* **156,** 229–234 (2014).
195. Li, Y. *et al.* Identification of citrullinated histone H3 as a potential serum protein biomarker in a lethal model of lipopolysaccharide-induced shock. *Surgery* **150,** 442–451 (2011).
196. Watson, C. T. *et al.* Integrative transcriptomic analysis reveals key drivers of acute peanut allergic reactions. *Nat Commun* **8,** 1943–13 (2017).
197. Jang, B. *et al.* Peptidylarginine deiminase and protein citrullination in prion diseases: strong evidence of neurodegeneration. *Prion* **7,** 42–46 (2013).
198. Ishigami, A. *et al.* Abnormal accumulation of citrullinated proteins catalyzed by peptidylarginine deiminase in hippocampal extracts from patients with Alzheimer's disease. *Journal of Neuroscience Research* **80,** 120–128 (2005).
199. Lange, S. *et al.* Protein deiminases: new players in the developmentally regulated loss of neural regenerative ability. *Dev. Biol.* **355,** 205–214 (2011).
200. Lazarus, R. C. *et al.* Protein Citrullination: A Proposed Mechanism for Pathology in Traumatic Brain Injury. *Front Neurol* **6,** 204 (2015).
201. Musse, A. A. *et al.* Peptidylarginine deiminase 2 (PAD2) overexpression in transgenic mice leads to myelin loss in the central nervous system. *Dis. Model. Mech.* **1,** 229–240 (2008).
202. Suzuki, A. *et al.* Decreased severity of experimental autoimmune arthritis in peptidylarginine deiminase type 4 knockout mice. *BMC Musculoskeletal Disorders* **17,** (2016).
203. Raijmakers, R. *et al.* Experimental autoimmune encephalomyelitis induction in peptidylarginine deiminase 2 knockout mice. *Journal of Comparative Neurology* **498,** 217–226 (2006).
204. Mastronardi, F. G. *et al.* Increased citrullination of histone H3 in multiple sclerosis brain and animal models of demyelination: a role for tumor necrosis factor-induced peptidylarginine deiminase 4 translocation. *J. Neurosci.* **26,** 11387–11396 (2006).
205. van Beers, J. J. B. C., Zendman, A. J. W., Raijmakers, R., Stammen-Vogelzangs, J. & Pruijn, G. J. M. Peptidylarginine deiminase expression and activity in PAD2 knock-out and PAD4-low mice. *Biochimie* **95,** 299–308 (2013).
206. Moscarello, M. A. *et al.* Inhibition of peptidyl-arginine deiminases reverses protein-hypercitrullination and disease in mouse models of multiple sclerosis. *Dis. Model. Mech.* **6,** 467–478 (2013).
207. Ligong Cao, Richard Goodin, Denise Wood, Mario A Moscarello, A.John N Whitaker. Rapid Release and Unusual Stability of Immunodominant Peptide 45−89 from Citrullinated Myelin Basic Protein†. *Biochemistry* **38,** 6157–6163 (1999).
208. Falcao, A. M. *et al.* PAD2-Mediated Citrullination Contributes to Efficient Oligodendrocyte Differentiation and Myelination. *Cell Rep* **27,** 1090–+ (2019).
209. Cao, L. *et al.* Inhibition of experimental allergic encephalomyelitis in the Lewis rat by paclitaxel. *J. Neuroimmunol.* **108,** 103–111 (2000).
210. Wei, L. *et al.* Novel Inhibitors of Protein Arginine Deiminase with Potential Activity in Multiple Sclerosis Animal Model. *J. Med. Chem.* **56,** 1715–1722 (2013).
211. Tejeda, E. J. C. *et al.* Noncovalent Protein Arginine Deiminase (PAD) Inhibitors Are Efficacious in Animal Models of Multiple Sclerosis. *J. Med. Chem.* **60,** 8876–8887 (2017).
212. Traka, M., Podojil, J. R., McCarthy, D. P., Miller, S. D. & Popko, B. Oligodendrocyte death results in immune-mediated CNS demyelination. *Nat. Neurosci.* **19,** 65–74 (2016).
213. Nicholas, A. P., Sambandam, T., Echols, J. D. & Tourtellotte, W. W. Increased citrullinated glial fibrillary acidic protein in secondary progressive multiple sclerosis.

Journal of Comparative Neurology **473,** 128–136 (2004).
214. Willis, V. C. *et al.* N-α-benzoyl-N5-(2-chloro-1-iminoethyl)-L-ornithine amide, a protein arginine deiminase inhibitor, reduces the severity of murine collagen-induced arthritis. *J. Immunol.* **186,** 4396–4404 (2011).
215. Seri, Y. *et al.* Peptidylarginine deiminase type 4 deficiency reduced arthritis severity in a glucose-6-phosphate isomerase-induced arthritis model. *Sci Rep* **5,** (2015).
216. Willis, V. C. *et al.* Protein arginine deiminase 4 inhibition is sufficient for the amelioration of collagen induced arthritis. *Clinical & Experimental Immunology* **188,** 263–274 (2017).
217. Rohrbach, A. S., Hemmers, S., Arandjelovic, S., Corr, M. & Mowen, K. A. PAD4 is not essential for disease in the K/BxN murine autoantibody-mediated model of arthritis. *Arthritis Res. Ther.* **14,** R104 (2012).
218. Suzuki, A. *et al.* Functional haplotypes of PADI4, encoding citrullinating enzyme peptidylarginine deiminase 4, are associated with rheumatoid arthritis. *Nature Genetics* **34,** 395–402 (2003).
219. Barton, A. *et al.* A functional haplotype of the PADI4 gene associated with rheumatoid arthritis in a Japanese population is not associated in a United Kingdom population. *Arthritis & Rheumatism* **50,** 1117–1121 (2004).
220. Martinez, A. *et al.* PADI4 polymorphisms are not associated with rheumatoid arthritis in the Spanish population. *Rheumatology* **44,** 1263–1266 (2005).
221. Caponi, L. *et al.* A family based study shows no association between rheumatoid arthritis and the PADI4 gene in a white French population. *Ann Rheum Dis* **64,** 587–593 (2005).
222. Kang, C. P. *et al.* A functional haplotype of the PADI4 gene associated with increased rheumatoid arthritis susceptibility in Koreans. *Arthritis & Rheumatism* **54,** 90–96 (2006).
223. Ikari, K. *et al.* Association between PADI4 and rheumatoid arthritis: A replication study. *Arthritis & Rheumatism* **52,** 3054–3057 (2005).
224. Plenge, R. M. *et al.* Replication of Putative Candidate-Gene Associations with Rheumatoid Arthritis in >4,000 Samples from North America and Sweden: Association of Susceptibility with PTPN22, CTLA4, and PADI4. *The American Journal of Human Genetics* **77,** 1044–1060 (2005).
225. Stahl, E. A. *et al.* Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nature Genetics* **42,** 508–514 (2010).
226. Okada, Y. *et al.* Meta-analysis identifies nine new loci associated with rheumatoid arthritis in the Japanese population. *Nature Genetics* **44,** 511 (2012).
227. Kurreeman, F. A. S. *et al.* Use of a Multiethnic Approach to Identify Rheumatoid-Arthritis-Susceptibility Loci, 1p36 and 17q12. *The American Journal of Human Genetics* **90,** 524–532 (2012).
228. Rantapaa-Dahlqvist, S. *et al.* Antibodies against cyclic citrullinated peptide and IgA rheumatoid factor predict the development of rheumatoid arthritis. *Arthritis & Rheumatism* **48,** 2741–2749 (2003).
229. Johansson, L. *et al.* Antibodies directed against endogenous and exogenous citrullinated antigens pre-date the onset of rheumatoid arthritis. *Arthritis Res. Ther.* **18,** –11 (2016).
230. Wipke, B. T. & Allen, P. M. Essential role of neutrophils in the initiation and progression of a murine model of rheumatoid arthritis. *J Immunol* **167,** 1601–1608 (2001).
231. Darrah, E. *et al.* Erosive Rheumatoid Arthritis Is Associated with Antibodies That Activate PAD4 by Increasing Calcium Sensitivity. *Science translational medicine* **5,** 186ra65–186ra65 (2013).
232. Turunen, S., Koivula, M.-K., Risteli, L. & Risteli, J. Anticitrulline antibodies can be caused by homocitrulline-containing proteins in rabbits. *Arthritis & Rheumatism* **62,** 3345–3352 (2010).
233. Shi, J. *et al.* Autoantibodies recognizing carbamylated proteins are present in sera of patients with rheumatoid arthritis and predict joint damage. *PNAS* **108,** 17372–17377 (2011).
234. Turunen, S. *et al.* Different amounts of protein-bound citrulline and homocitrulline in

foot joint tissues of a patient with anti-citrullinated protein antibody positive erosive rheumatoid arthritis. *J Transl Med* **11,** 224 (2013).
235. Turunen, S., Hannonen, P., Koivula, M.-K., Risteli, L. & Risteli, J. Separate and overlapping specificities in rheumatoid arthritis antibodies binding to citrulline- and homocitrulline-containing peptides related to type I and II collagen telopeptides. *Arthritis Res. Ther.* **17,** 2 (2015).
236. Horibata, S. *et al.* Role of peptidylarginine deiminase 2 (PAD2) in mammary carcinoma cell migration. *BMC Cancer* **17,** 378 (2017).
237. Wang, L. *et al.* PADI2-Mediated Citrullination Promotes Prostate Cancer Progression. *Cancer Res.* **77,** 5755–5768 (2017).
238. Cantariño, N. *et al.* Downregulation of the Deiminase PADI2 Is an Early Event in Colorectal Carcinogenesis and Indicates Poor Prognosis. *Mol. Cancer Res.* **14,** 841–848 (2016).
239. Yuzhalin, A. E. *et al.* Colorectal cancer liver metastatic growth depends on PAD4-driven citrullination of the extracellular matrix. *Nat Commun* **9,** 4783 (2018).
240. McNee, G. *et al.* Citrullination of histone H3 drives IL-6 production by bone marrow mesenchymal stem cells in MGUS and multiple myeloma. *Leukemia* **31,** 373–381 (2017).
241. Nakashima, K. *et al.* PAD4 regulates proliferation of multipotent haematopoietic cells by controlling c-myc expression. *Nat Commun* **4,** (2013).
242. Chang, X. & Han, J. Expression of peptidylarginine deiminase type 4 (PAD4) in various tumors. *Molecular Carcinogenesis* **45,** 183–196 (2006).
243. Chang, X. *et al.* Increased PADI4 expression in blood and tissues of patients with malignant tumors. *BMC Cancer* **9,** 40 (2009).
244. Tanikawa, C. *et al.* Regulation of histone modification and chromatin structure by the p53–PADI4 pathway. *Nat Commun* **3,** 676 (2012).
245. Robbins, P. F. *et al.* Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells. *Nat Med* **19,** 747–752 (2013).
246. Linnemann, C. *et al.* High-throughput epitope discovery reveals frequent recognition of neo-antigens by CD4⁺ T cells in human melanoma. *Nat Med* **21,** 81–85 (2015).
247. Ireland, J. M. & Unanue, E. R. Autophagy in antigen-presenting cells results in presentation of citrullinated peptides to CD4 T cells. *J Exp Med* **208,** 2625–2632 (2011).
248. Brentville, V. A. *et al.* Citrullinated Vimentin Presented on MHC-II in Tumor Cells Is a Target for CD4(+) T-Cell-Mediated Antitumor Immunity. *Cancer Res.* **76,** 548–560 (2016).
249. Durrant, L. G., Metheringham, R. L. & Brentville, V. A. Autophagy, citrullination and cancer. *Autophagy* **12,** 1055–1056 (2016).
250. Cook, K. *et al.* Citrullinated alpha-enolase is an effective target for anti-cancer immunity. *Oncoimmunology* **7,** (2018).
251. Kolodziej, S. *et al.* PADI4 acts as a coactivator of Tal1 by counteracting repressive histone arginine methylation. *Nat Commun* **5,** (2014).
252. Krivtsov, A. V. *et al.* Transformation from committed progenitor to leukaemia stem cell initiated by MLL–AF9. *Nature* **442,** 818–822 (2006).
253. Bicker, K. L. & Thompson, P. R. The protein arginine deiminases: Structure, function, inhibition, and disease. *Biopolymers* **99,** 155–163 (2013).
254. Subramanian, V., Slade, D. J. & Thompson, P. R. in *Protein Deimination in Human Health and Disease* 377–427 (Springer New York, 2014). doi:10.1007/978-1-4614-8317-5_21
255. Pritzker, L. B. & Moscarello, M. A. A novel microtubule independent effect of paclitaxel: the inhibition of peptidylarginine deiminase from bovine brain. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology* **1388,** 154–160 (1998).
256. Knuckley, B., Luo, Y. & Thompson, P. R. Profiling Protein Arginine Deiminase 4 (PAD4): a novel screen to identify PAD4 inhibitors. *Bioorg. Med. Chem.* **16,** 739–745 (2008).

257. Drenth, J., Kalk, K. H. & Swen, H. M. Binding of Chloromethyl Ketone Substrate Analogs to Crystalline Papain. *Biochemistry* **15,** 3731–3738 (1976).
258. Kreutter, K. *et al.* Three-dimensional structure of chymotrypsin inactivated with (2S)-N-acetyl-L-alanyl-L-phenylalanyl alpha-chloroethane: implications for the mechanism of inactivation of serine proteases by chloroketones. *Biochemistry* **33,** 13792–13800 (1994).
259. Powers, J. C., Asgian, J. L., Ekici, O. D. & James, K. E. Irreversible inhibitors of serine, cysteine, and threonine proteases. *Chem. Rev.* **102,** 4639–4750 (2002).
260. Luo, Y., Knuckley, B., Lee, Y.-H., Stallcup, M. R. & Thompson, P. R. A fluoroacetamidine-based inactivator of protein arginine deiminase 4: design, synthesis, and in vitro and in vivo evaluation. *J. Am. Chem. Soc.* **128,** 1092–1093 (2006).
261. Luo, Y. *et al.* Inhibitors and inactivators of protein arginine deiminase 4: functional and structural characterization. *Biochemistry* **45,** 11727–11736 (2006).
262. Wang, Y. *et al.* Anticancer Peptidylarginine Deiminase (PAD) Inhibitors Regulate the Autophagy Flux and the Mammalian Target of Rapamycin Complex 1 Activity. *J. Biol. Chem.* **287,** 25941–25953 (2012).
263. Knight, J. S. *et al.* Peptidylarginine deiminase inhibition disrupts NET formation and protects against kidney, skin and vascular disease in lupus-prone MRL/lpr mice. *Ann Rheum Dis* **74,** 2199–2206 (2015).
264. Lewis, H. D. *et al.* Inhibition of PAD4 activity is sufficient to disrupt mouse and human NET formation. *Nat. Chem. Biol.* **11,** 189–191 (2015).

Chapter 2: Materials and Methods

2.1 Biochemical methods

2.1.1 Bacterial transformations

Chemically competent DH5α Escherichia coli (E. coli) (sub-cloning efficiency, Invitrogen) were used for production of DNA and BL21 DE3 cells (New England BioLabs) for the production of recombinant proteins. Competent E. coli were thawed on ice for 10 min from -80˚C and pipetted into 50 µL aliquots in pre cooled sterile Eppendorf tubes. An appropriate amount of plasmid (1 pg–100 ng) was added to each tube, which was flicked 4–5 times to mix and incubated on ice for 30 min. Bacteria were heat shocked at 42°C for 45 sec, placed on ice for 2 min, and 450 µl room temperature SOC media (MRC Human Genetics Unit (HGU) Technical Services) was added for outgrowth at 37˚C with shaking for 1 h. Cells were mixed by inverting and two different amounts (50 µL, 450 µL) were spread onto L-agar plates supplemented with the appropriate antibiotic (final concentrations of ampicillin at 100 µg/mL; kanamycin at 50 µg/mL; chloramphenicol at 13 µg/mL; L-agar plates coated with the appropriate antibiotic were prepared by MRC HGU technical services). Plates were inverted and incubated at 37°C overnight to allow for colony growth.

2.1.2 Starter cultures and plasmid preparation

Single colonies were picked in the morning using a sterile toothpick and inoculated into 3 mL Luria-Bertani (LB) broth (MRC HGU Technical Services) in Falcon Round-Bottom SnapCap tubes supplemented with the appropriate concentration of antibiotic. These were grown at 37˚C with shaking at 300 rpm until the afternoon and 500 µL used to inoculate 50 mL of LB media cultures, which were grown overnight at 37˚C with shaking at 300 rpm in a New Brunswick Scientific G25 environmental shaker. The next day pellets were harvested at 4500 rpm for 20 min. Qiagen Maxiprep kit was used to extract and purify plasmids according to manufacturer's instructions, and

DNA was eluted into TE buffer, pH 8.0 or sterile MilliQ water (buffers prepared by MRC HGU Technical Services).

2.1.3 Obtaining DNA sequences

Sequences for the putative peptidyl arginine deiminase from *Cyanothece sp. 8801* and *Porphyrmonas gingivalis* were obtained from National Center for Biotechnology Information (NCBI) databases. Sequences were designed to be flanked by EcoRI (at the 5' end) and XhoI (at the 3' end) restriction sites and were synthesized by ThermoFisher Scientific GeneArt. Vectors for PADI2beta, PADI2 and PADI3 were a generous gift from Dr Ana Mendanha Falcao and Prof Castelo Branco, Karolinska Institutet. Other PADI sequences were obtained from Dr Christophorou. All sequences were subcloned using InFusion cloning methods into a modified pGEX 6P-1 bacterial expression vector which contains an additional 6xHis tag included C-terminally to the glutathione S-transferase (GST) tag sequence, at the N-terminus of the target protein (generous gift from Dr Martin Reijns, MRC Human Genetics Unit). InFusion subcloning was performed by Gavriil Gavrilidis and Abigail Wilson.

2.1.4 Glycerol stocks

To make glycerol stocks, 500 µL of an overnight bacterial culture was added to 500 µL of 50% glycerol in dd H_2O in a 2 mL Cryogenic Vial and gently mixed before freezing at -80˚C. Bacteria were recovered from the glycerol stock by scraping frozen bacteria with a sterile loop from the cryovial without allowing the glycerol stock to thaw. Cells were streaked on to agar plates with the appropriate antibiotic. Plates were inverted and incubated at 37˚C overnight to grow colonies, which were picked to prepare starter cultures for further propagation.

2.1.5 Recombinant protein production

A single colony was picked from an agar plate to inoculate a starter culture flask of 30 mL 2TY media (MRC HGU Technical Services) containing the

appropriate antibiotic in the evening. This was grown overnight at 37˚C with shaking at 220 rpm. The next morning, an empty incubator was prepared with 2L flask adaptors and set to 18˚C. For each protein, 3x 2L culture flasks were prepared with 400 mL of 2TY media per flask and appropriate antibiotic. 4mL (1:100 dilution) of mixed starter culture bacteria was used to inoculate each large 2L flask. These were grown at 37˚C for approximately 2 hours with shaking at 220 rpm. Optical density at a wavelength of 600 nm (OD600) was measured at 1 hour and then regularly after that. At OD600=0.6, cultures were induced with isopropyl β-D-1-thiogalactopyranoside (IPTG) (Sigma) at a final concentration of 0.5 mM and flasks were transferred to the pre-cooled 18˚C shaker. These were incubated overnight with shaking at 220 rpm. Bacteria were pelleted at 8000 x g at 4˚C for 15 min, weighed and frozen at -80˚C.

2.1.6 Protease and Phosphatase inhibitors

For protease inhibitors, a Roche cOmplete™, Mini, ethylenediaminetetraacetic acid (EDTA)-free Protease Inhibitor Cocktail tablet was dissolved in the buffer freshly. For phosphatase inhibitors Roche Phos-STOP inhibitor tablet was dissolved in the buffer freshly. For large buffer volumes, protease inhibition was achieved using 1mM PMSF.

2.1.7 Fast protein liquid chromatography (FPLC)

Bacterial pellets were thawed on ice, transferred into 50mL falcon tubes and lysis buffer was added at 4mL per g of bacterial pellet. Lysis buffer comprises binding buffer supplemented with 1 mM (final) dithiothreitol (DTT) (Sigma), 2 mM (final) $MgCl_2$ (Sigma) and benzonase (1 µL) (Sigma). Falcons were rotated at 4˚C to dissolve the pellet for 20 min, sonicated with a Soniprep 150 for 7 cycles, 45 sec on, 45 sec off at amplitude 8. Lysates were cleared by centrifugation at 20 000 x g for 20 min and the supernatant passed through a polyethersulfone (PES) 0.22 µm syringe filter (Millipore) before loading onto an fast protein liquid chromatography (FPLC) machine (AKTA). Purification was carried out using a 5 mL nickel affinity His-Trap

column. The column was cleaned before equilibration in binding buffer for 5 column volumes (CV). Lysate was loaded via Superloop, and the column washed in Wash buffer for 5 CV, before elution over 5 CV in Elution buffer in 1 mL fractions. Fractions were analysed by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE), pooled and concentrated by molecular weight cutoff filter (Vivaspin) to approximately 5-10 mg/mL into storage buffer before flash freezing.

FPLC buffers	NaCl	Tris-HCl	Imidazole	Additives
Binding buffer	500 mM	50 mM	20mM	pH~7.5, 5% Glycerol, 1mM DTT
Wash buffer	500 mM	50 mM	25mM	pH~7.5, 5% Glycerol, 1mM DTT
Elution buffer	500 mM	50 mM	250mM	pH~7.5, 5% Glycerol, 1mM DTT
Storage buffer	150mM	50mM	–	pH~7.5, 5% Glycerol, 1mM DTT

Figure 2.1: Table of FPLC buffers

2.1.8 Protein concentration determination

Protein concentrations were determined using 1 µL protein solution on a NanoDrop 8000 UV/Vis spectrophotometer with absorbance measured at 280 nm. ε was obtained using the ProtParam tool on the ExPASY Bioinformatics Resource Portal from the Swiss Institute of Bioinformatics (https://web.expasy.org/protparam/) and concentrations calculated with the Beer Lambert law (A = ε.c.l).

2.1.9 Optical density measurements

OD600 values were measured in a 1.6 mL cuvette, against a reference sample of initial growth medium using a spectrophotometer.

2.1.10 Whole cell lysate extraction

Media was removed and cells washed once in ice-cold phosphate buffered saline (PBS). 100 µL of modified 2x Laemmli (120 mM Tris-HCl, 20% glycerol, 4% SDS, pH 6.8) was added to each well and scraped using a silicon cell scraper (Corning). The extract was transferred to a 1.5 mL Eppendorf tube, boiled at 95˚C for 5 min and spun quickly. Samples were passed through a 25G needle ten times, sonicated for 5 min using a Bioruptor® water bath sonicator at high power in 30-second cycles or sonicated for 3 x 30 secs using an Ultrasonic disintegrator Soniprep 150. Samples were sometimes stored at –80°C prior to normalization. Total protein concentration was determined using the nanodrop and concentrations were equalised in 2 x Laemmli buffer. Equalised protein mixtures were added 9:1 to a pre-prepared mixture of 10% bromophenol blue in 98% β-mercaptoethanol. Samples were then boiled for 5 min and spun quickly and loaded for SDS PAGE gel running or stored at –80˚C.

2.1.11 SDS-PAGE running

Samples were heated at 95˚C for 5 min and spun quickly before loading. Polyacrylamide gels were run for 40 min at 200 V in running buffer using either a Bio-Rad Mini-PROTEAN II or CRITERION system with 4-20% TGX Stain-Free pre-cast gels. 30 µg of total protein was loaded per well as standard using SDS-PAGE. Proteins were stained with Coomassie Blue stain or subjected to Western Blotting.

2.1.12 Coomassie staining

Coomassie staining was performed by incubating in Coomassie stain (Acetic acid 10% (w/v), 30% Methanol in 0.25% (w/v) Coomassie® brilliant blue) for 30 min before washing in Destain solution (Acetic acid 10% (w/v), 30% Methanol (w/v)) several times up to overnight and washing once in MilliQ water.

2.1.13 Western blotting

After SDS-PAGE running, proteins were transferred for 60 min at 400 mA using either a Bio-Rad Mini-PROTEAN II or CRITERION system in transfer buffer onto nitrocellulose membrane (0.45 µM, BioRad) with an ice pack included. Membranes were blocked in 5% bovine serum albumin (BSA) in Tris-buffered saline (TBS) with 0.5% Tween-20 detergent added (TBS-T) for 1 h at room temperature. Membranes were incubated in an appropriate dilution of primary antibody in blocking buffer, with agitation overnight at 4°C. Membranes were washed in a minimum of three washes of TBS-T: 5 min/20 min/5 min, before incubation in secondary antibody diluted in blocking buffer with agitation at room temperature for 1 hour. Membranes were washed again as above. For signal development, Thermo Scientific™ SuperSignal™ West Pico PLUS Chemiluminescent Substrate was used as per manufacturer's instructions. Excess reagent was removed and membranes placed in transparent plastic before image acquisition using a GE ImageQuant LAS 4000.

2.1.14 SDS-PAGE and Western blot buffers:

Name:	Tris	Glycine	Detergent	Protein	pH	Solvent
Running buffer	25 mM	190 mM	0.1% SDS		pH = 8.3	H_2O
Transfer buffer	25 mM	190 mM			pH = 8.3	20% MeOH, 80% H_2O
TBS-T	1xTBS		0.5% (v/v) Tween-20			
TBS-lo-T	1xTBS		0.1% (v/v) Tween-20			
Blocking	1xTBS		0.5% (v/v) Tween-20	5% (w/v) BSA		

Figure 2.2: Table of SDS-PAGE and Western blot buffers

2.1.15 Antibody use for Western blotting

The following commercial antibodies were used at these stated dilutions: GAPDH ab127428 GR259405-4 (1:10000), NPM1 ab37659 lot# GR187575-1 (1:500), H3CitR2 ab212082 monoclonal (1:2000), PADI4 ab50332 GR291996-8/291996-1 (1:2000), total histone H3 ab10799 (1:1000), anti-calmodulin EP799Y ab45689 (1:1000), anti-GST Glutathione S-Transferase ab19256 (1:1000), anti-mouse Padi2 ab50257 (1:1000), abcam anti-6xHis ab18184 (1:1000) from Abcam; non-phosphorylated (Active) β-Catenin (Ser45) (D2U8Y) XP®Rabbit mAb #19807 (1:1000), non-phosphorylated (Active) β-Catenin (Ser33/37/Thr41) (D13A1) Rabbit mAb #8814 (1:1000), total β-Catenin (D10A8) XP® Rabbit mAb #8480 (1:1000), GSK-3β (D5C5Z) XP® Rabbit mAb #12456 (1:1000) from Cell Signaling Technologies; and total β-Catenin (1:1000) from BD Millipore. The polyclonal H3CitR(2,8,17) ab5103 antibody is known to show lot to lot variability. A number of different lots were tested and validated using activation conditions in HL60 cells and only the following two lots were used: GR99830-1 (1:2000) and GR247556-1 (1:5000).

2.1.16 Mod-Cit Western blotting

100 µg total cell lysate was loaded per lane and SDS-PAGE run for 40 min at 200 V. Proteins were then transferred onto nitrocellulose wetted in transfer buffer for 1 hour at 400 mA. 10 mL reagent A was made fresh before use per two membranes during transfer. Reagent A was prepared by diluting 0.5 mL 0.5% $FeCl_3$ in 5 mL MiliQ water on ice, before dropwise addition with constant mixing of 2.5 mL 98% H_2SO_4 and 2 mL 85% H_3PO_4 in a fume hood and leaving to cool at 4˚C (Sigma). After transfer, membranes were washed twice in MilliQ, blocked with 0.1% ovalbumin in MilliQ for 15 min at RT with agitation and washed twice again in MilliQ. Reagent B (0.5% diacetyl monoxime, 0.25% antipyrine, and 0.50 mol/L acetic acid) was prepared in advance and stored frozen at -20˚C (Sigma). Reagent A and B were thawed and mixed 1:1 and incubated with the membrane overnight, protected from light without agitation or CO_2. The final concentration for the chemical

modification mixture was 0.0125% $FeCl_3$, 2.3 mol/L H_2SO_4, 1.5 mol/L H_3PO_4, 0.25% diacetyl monoxime, 0.125% antipyrine, and 0.25 mol/L acetic acid. Membranes were then washed 4-5 times in MilliQ water, blocked in 3% milk in TBS with 0.1% Tween-20 detergent added (TBS-lo-T) for 1h at RT, and incubated with anti-ModCit primary antibody at a 1:500 dilution for 3 h at RT followed by incubation at 4˚C overnight. Membranes were then washed three times with TBS-T. Secondary antibody incubation was with goat anti-Rabbit secondary at a 1:2,000 dilution in TBS-T for 3 h at RT. Standard detection procedures as above were used.

2.1.17 Nuclear and cytoplasmic extraction

The Thermo Scientific Pierce NE-PER Nuclear and Cytoplasmic Extraction kit was used for stepwise separation and preparation of cytoplasmic and nuclear extracts from cultured cells according to manufacturer's recommendations.

2.1.18 Cyclic peptides

Lyophilised cyclic peptides, synthesized by Dr. Louise Walport (University of Tokyo), were initially reconstituted in 20 µL dimethyl sulfoxide (DMSO) with additional 15 µL added increasingly until the solid peptide could be dissolved. 0.5 µL was diluted in 2.5 µL DMSO before measuring concentrations using the Nanodrop 8000 UV/Vis spectrophotometer with absorbance measured at 280 nm. Extinction coefficients (ε) were obtained using the ProtParam tool on the ExPASY Bioinformatics Resource Portal from the Swiss Institute for Bioinformatics (https://web.expasy.org/protparam/) and concentrations calculated with the Beer Lambert law (A = ε.c.l). For biotinylated peptides ε was given to be 4470 $M^{-1}cm^{-1}$. Fluorescent peptides were measured at 495 nm instead of 280 nm, using the ε of the fluorescein amidite (FAM) fluorophore of 76900 $M^{-1}cm^{-1}$. 10 mM stocks of peptide were prepared, stored at -20˚C and protected from light.

2.2 Cell biological methods

All procedures were carried out in a Class II Biological Safety Cabinet.

2.2.1 Mouse ES cell culture

E14 embryonic stem (ES) cells were cultured in serum media following *Christophorou et al.* 2014[1], but without pencillin/streptomycin. To passage cells, medium was aspirated and cells washed gently in PBS. 3 mL of pre-warmed Accutase® was incubated over the surface of the flask for 2 – 2.5 min until cells detached and 9 mL of full serum medium added to inactivate the Accutase®. Cells were centrifuged at 1200 rpm for 4 min and medium was aspirated. The cells were resuspended and triturated ten times in 1 mL of full serum medium using a P1000 pipette to break up clumps of cells remaining in colonies. Cells were counted using a haemocytometer and reseeded in pre-warmed full serum medium (see below for seeding densities). Medium was replaced daily and cells not allowed to attain greater than 80% confluency. Health, colony formation and confluency were assessed at each media change. To freeze cells, 2-3 million cells per 500 µL full serum medium was added to 500 µL of Cryomix (full serum media containing 20% DMSO and 20% fetal calf serum (FCS), MRC HGU Technical Services). Cells were mixed by gentle pipetting and frozen in a Mr Frosty box at -80˚C overnight before transfer to liquid nitrogen storage. For reviving cells, vials were removed from liquid nitrogen and kept under dry ice until thawing. A single 1 mL vial (2-3 million cells as above) was thawed rapidly at 37˚C in a water bath and added to 9 mL pre-warmed medium to dilute the DMSO content. Cells were centrifuged at 1200 rpm for 4 min, media aspirated and cells resuspended in 5 mL full serum medium. These were seeded in a T25 flask containing 6 mL medium. Medium was changed daily and cells split having attained 70-80% confluency and passaged thereafter as above.

Size of flash	Number of cells at seeding
T75 Flask	2-3 million
T25 Flask	1-1.5 million
6-Well Plate	3-400 000 per well
24-Well Plate	100 000 per well
96-well Plate	10 000 per well

Figure 2.3: Table of mES cell seeding densities

2.2.1.1 Mouse ES cell media (Serum media)

Glasgow's modified Eagle's Minimal Essential Medium (GMEM) BHK21 (ThermoFisher) was prepared with 10% FCS (MRC HGU Technical Services), 1X Non-Essential Amino acids (Sigma 100X stock), 1 mM Sodium Pyruvate (Sigma 100 mM stock), 2 mM L-Glutamine (200 mM stock from MRC HGU Technical Services), 100 µM β-mercaptoethanol (50 mM Gibco stock) and recombinant leukaemia inhibitory factor (LIF) (either 1:500 from homemade stock from Abigail Wilson, or 10^3 units from ESGRO® Recombinant Mouse LIF Protein).

2.2.1.2 Mouse ES cell media (KSR media)

GMEM BHK21 medium, 1% FCS, 10% Gibco KnockOut Serum Replacement (KSR), 1X Non-Essential Amino acids (Sigma 100X stock), 1 mM Sodium Pyruvate (Sigma 100 mM stock), 2 mM L-Glutamine (200 mM stock from MRC HGU Technical Services), 100 µM β-mercaptoethanol (50 mM Gibco stock), recombinant LIF (either 1:500 from homemade stock from Abigail Wilson, or 10^3 units from ESGRO® Recombinant Mouse LIF Protein).

2.2.2 Pre-iPS cell culturing and reprogramming assay

Pre-induced pluripotent stem cells (pre-iPS) cells are derived from neural stem cells (NSO4G) that have been tagged at the endogenous Oct4 locus with green fluorescent protein (GFP). NSO4G cells are transduced with three Yamanaka factors (human Oct4, Klf4 and c-Myc) since neural stem cells already express the fourth Yamanaka factor Sox2 at day 0. These cells are

then cultured in neural stem cell media for 4 days and subsequently in mouse ES cell media supplemented with recombinant LIF (10^3 units, ESGRO recombinant murine LIF) for 2 days before freezing following *Theunissen et al.* or subsequently *Christophorou et al.*[1,2]. These cells (i.e. before addition of 2i media at day 6 of a reprogramming experiment) are referred to as "pre-iPS cells". As cells reprogram, they begin to express the endogenous murine Oct4, in place of the exogenous transduced Oct4, and therefore express GFP. A normal successful reprogramming experiment cultures pre-iPS cells in KSR2i media (KSR media supplemented with 1μM PD0325901 AxonMedChem and 3 μM CHIR99021 AxonMedChem) for an additional 8 days and achieves ~5% reprogramming efficiency by day 14 as read out by the proportion of GFP positive cells. Culturing in full serum media, or in Cl-amidine reduces the efficiency of reprogramming after *Christophorou et al.*[1].

2.2.3 HL-60 cell culturing

HL-60 promyelocytic leukaemia cells (ATCC® CCL-240™) were maintained in Roswell Park Memorial Institute 1640 medium (RPMI) (ThermoFisher) was supplemented with 2 mM L-Glutamine and 10% FCS (MRC HGU Technical Services). Cultures were passaged by centrifugation at 1000 rpm for 5 min with subsequent resuspension at a density of $1x10^5$ viable cells per mL in Corning® T-75 flasks (catalog #431464). Cell concentrations were not allowed to exceed $1x10^6$ cells per mL. Medium was renewed every 2 to 3 days. Terminal differentiation was induced by culturing in media supplemented with 1.25% DMSO (Hybri-Max, Sigma) over 5 days after which robust PADI4 expression can be detected by Western blotting[3]. The morphology of the cells changes to become smaller and more densely coloured (less transparent- presumably due to granule formation) under a microscope and the cells no longer divide.

2.2.4 Neutrophil isolation from peripheral blood and short-term culture

The protocol for isolation of primary human neutrophils was adapted with initial instruction of Dr David Dorward from the lab of Prof Adriano Rossi,

MRC Centre for Inflammation Research, Queen's Medical Research Institute (QMRI) and then subsequently performed at the Institute for Genetics and Molecular Medicine (IGMM). All procedures were carried out at room temperature unless otherwise stated. All centrifugation steps were performed with capped rotor buckets. Falcon brand 15 and 50 mL tubes were used for the preparation. 3.8% sodium citrate solution was prepared in advance from 3.8 g tribasic dihydrate (Sigma #25116) dissolved in 100 mL Baxter UKF7114 sterile water. 6% Dextran solution in saline was prepared in advance from 6 g Dextran 500 (Sigma #31392) dissolved in 100 mL warmed 0.9% NaCl solution (Baxter). All solutions were then sterile filtered using a 0.22 μm PES filter unit. Percoll was obtained from GE Healthcare #17-0891-02. 1x and 10x PBS used throughout were without Ca^{2+}/Mg^{2+} (Sigma).

Freshly drawn blood from healthy adult volunteers (presenting without hay fever due to high eosinophil populations that co-purify in this protocol with neutrophils) was collected either at the QMRI or Western General Hospital directly into 3.8% sodium citrate solution (4 mL of 3.8% citrate per 40 mL of blood in a 50 mL falcon tube) which was mixed by gentle inversion of tube with parafilm covering the cap. Blood was centrifuged at 350 g for 20 min (minimum or zero brake was used: Acc1/ Brake 0 Hettich centrifuge, Acc0/Brake 0 Mistral centrifuge). After centrifugation platelet-rich plasma (PRP) layer was aspirated without disturbing the pelleted cells. Autologous recalcified plasma (autologous serum) was prepared by adding 220 μl of 1 M CaCl2 per 10 mL plasma in glass tubes and incubated at 37°C for 1 hour. Autologous serum was separated from the platelet plug and transferred to a Falcon tube for use on the same day, but can be frozen and stored at -20˚C. Leukocytes were separated from erythrocytes by Dextran sedimentation: 6 mL of 6% Dextran in saline solution was added to each tube (2.5 mL Dextran per 10 mL cell pellet) and then made up to 50 mL using pre-warmed saline at 37°C. Cells were fully resuspended by careful, gentle mixing before allowing to sediment through the dextran for between 20 and 30 min (not more than 30 min) at room temperature by gravity. During this period, discontinuous

Percoll gradients were prepared. Room temperature stock Percoll solution was made isotonic using 10x PBS to yield a 90% “stock” solution by mixing 27 mL Percoll with 3 mL 10x PBS. The 81% Percoll layer solution was made with 8.1 mL Percoll “stock” and 1.9 mL 1xPBS; the 70% Percoll solution using 7 mL Percoll “stock” and 3 mL 1xPBS; and the 55% Percoll solution with 5.5 mL Percoll “stock” and 4.5 mL 1xPBS. A discontinuous gradient was prepared by placing 3 mL of 81% Percoll (bottom layer) at the bottom of a 15 mL Falcon tube with 3 mL of 70% Percoll (middle layer) overlayered onto the bottom layer slowly to avoid any mixing of the two different percentage Percoll layers. After dextran sedimentation, the leukocyte-rich upper layer was retained and transferred to a fresh 50 mL tube and topped up to 50 mL with saline and then centrifuged at 350g for 6 min (Acc5/ Brake 5 for both Hettich and Mistral centrifuges). Pellets from 2 x 50 mL Falcons were then resuspended in a combined 3 mL of 55% Percoll solution and overlayed as the final top layer onto the previously prepared Percoll gradient with care as before not to disturb the layers. Completed gradients were centrifuged at 720g for 20 min, (Acc1/ Brake 0 Hettich centrifuge, Acc0/Brake 0 Mistral centrifuge). Leukocytes were harvested at the 70%/81% interface (residual erythrocytes should pellet at the bottom of the Falcon and mononuclear cells can be harvested at the 55%/70% interface). Harvested leukocytes were then washed twice in PBS with centrifugation at 230 x *g* for 6 min, (Acc5/ Brake5 Hettich and Mistral centrifuges). Cells were analysed by cytospin and also by flow purity using Fluorescence-Activated Cell Sorting (FACS) with CD16 and CD14 mouse antibodies (mAbs) to ensure neutrophils (CD16 positive cells) represented at least 95% of the cell population. Cells were finally resuspended in appropriate buffer for short-term cell stimulation for PADI4 activation. Diff-Quick (Gamidor, Oxford, UK) was used to stain neutrophils prepared by cytospin (neutrophils have distinctive lobular nuclear morphology; eosinophils have distinctive orange granules).

2.2.5 Cell culture treatments

Cell culture treatments were performed as stated in experimental figure legends in Iscove's Modified Dulbecco's Medium (IMDM), GMEM, ES Serum media, ES KSR media, in Locke's solution (10 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES), pH7.3, 150 mM NaCl, 5 mM KCl, 2 mM $CaCl_2$, 0.1% glucose). Materials for cellular treatments were recombinant TNF alpha (human, suitable for cell culture, Sigma), LPS (Sigma), Interleukin-8 (IL-8) recombinant from *E. coli* (human, Sigma), N-Formyl-Met-Leu-Phe chemotactic peptide (fMLF) (Sigma), calcium ionophore A23187 (Sigma), all-trans-retinoic acid (ATRA) (cell culture, Sigma), DMSO (Hybrid-Max, Sigma), PD0325091 (AxonMedChem), CHIR99021 (AxonMedChem), Cl-amidine (Cayman Chemical), BB-Cl-amidine (Cayman Chemical), alpha amanitin (generous gift from Dr Ryu-Suke Nozawa, MRC HGU, Sigma), actinomycin D (Sigma), mouse Wnt3a recombinant (Peprotech), BML284 (Cayman Chemical).

2.2.6 Cellular Methods

2.2.6.1 Citrullination Lysate Assay

Mouse ES cells were cultured for 1-2 days in serum media on six well plates pre-coated with gelatin. Cells were washed in 1 mL PBS supplemented with 2 mM EDTA (MRC HGU Technical Services) and once in PBS. 100 µL of active lysis buffer (20 mM Tris, 1% NP-40, 5 mM DTT, 137 mM NaCl with protease inhibitors) was added to each well. Cells were scraped with a silicon scraper and combined into a single Eppendorf. 0.2 µL of 1 M $MgCl_2$ (2 mM $MgCl_2$ final) and 0.2 µL benzonase (Sigma) per 100 µL of lysate was added and incubated at 4˚C with rotation for 30 min. Sample was sheared with 25G needle 10 times and centrifuged at 10000 rpm at 4˚C for 5 min to clear cell debris and supernatant transferred to a new clean tube on ice. The overall lysate for the total number of conditions was pooled and split into each assay condition. Any supplements ($CaCl_2$, EDTA, recombinant bovine Calmodulin, peptide, Cl-amidine, GSK484, vehicle) were added at stock concentration in a single droplet to the tube lid before the assay was initiated. To begin the

assay tubes were spun quickly and incubated at the desired temperature and time with shaking in an Eppendorf heating block. To quench the assay samples were boiled at 95˚C for 5 min before freezing directly on dry ice.

2.2.6.2 Transfections

Transfections were performed using Lipofectamine 2000 (Invitrogen). The procedure is written here for mouse ES cells seeded in six well plates, but was scaled up according to manufacturer's instructions. In the morning, mouse ES cells were seeded in six well plates such that, by the next day, cells have attained approximately 50% confluency in colonies. Per well of transfection, 4 µg of plasmid DNA was mixed with 250 µL of pre-warmed OptiMEM medium (Gibco) in an Eppendorf. In a second Eppendorf, 10 µL Lipofectamine 2000 was mixed with 250µL OptiMEM (ThermoFisher) and incubated for 5 min at RT. The Lipofectamine 2000 mixture was then added to the first Eppendorf containing plasmid DNA without pipetting, and mixed by flicking and inverting the tube. This was then incubated for 20 min at RT. Cells to be transfected were washed once with PBS and incubated with 2 mL of plain GMEM (ThermoFisher) supplemented only with recombinant LIF (1:500 homemade, or 10^3 units of ESGRO recombinant murine LIF). After incubation has finished, the Lipofectamine DNA mixture was added dropwise to the cell while rocking the plate to mix and returned to the incubator. At four to six hours after transfection, fresh medium was replaced on the cells.

The Pierce™ Protein Transfection Reagent (ThermoFisher) was used according to manufacturer's instructions and tested using FAM tagged peptides 1, 2 and 3. These pilot experiments were not continued once activity was clear from direct usage on cells.

2.2.6.3 Nucleofection and generation of stable cell lines

Nucleofections were performed using the Lonza 4D-NucleofectorTM System according to manufacturer's instructions. pB-CAG-Ctrl or pB-CAG-hPADI4 vectors (1 mg) had been transfected with piggyBac transposase (pPBase)

expression vector, pCAG-PBase (2 mg), by nucleofection according to the manufacturer's instructions (Lonza) and as described in *Christophorou et al.*[1]. ES E14 cells constitutively expressing the hygromycin resistance gene and human PADI4 were selected and expanded in media containing 200 mg/mL hygromycin (Sigma). Other vectors introduced using the same method were piggyBac-AVI-PADI2b-IRES-GFP, piggyBac-AVI-PADI3-IRES-GFP, piggyBac-PADI2-Hygromycin, and piggyBac-BirA-Hygromycin. In detail, a T25 flask of cells was incubated with Accutase to detach cells, centrifuged at 1200 rpm for 4 min and resuspended in 5 mL volume media after triturating in a 1 mL volume. Cells were then put in an uncoated T25 flash to allow cells to recover for 20 min in the incubator. Cells were counted and $5x10^4$cells per sample were collected in a 15 mL Falcon tube and centrifuged as above to pellet. An extra mock condition for nucleofection without plasmid and a final untreated condition for assessing death kinetics under hygromycin treatment were always included. Each sample is resuspended in 20 µL of Nucleofector solution. Plasmids of 0.3 µg pB-CAG-Vector and 0.6 µg pCAG-PBase were added to a 1.5 mL Eppendorf per sample, prepared using EtOH sprayed pipettes and filter tips. The 20 µL cell suspension in Nucleofector solution was added to the DNA, mixed and transferred to the 16 well Amaxa cuvette ensuring there were no air bubbles. The CG-104 (ES cell) program was used to nucleofect a 16-well cuvette. Cell suspension was then made up to 2 mL volume of pre-warmed media and seeded onto a six well plate. After 24 hours of recovery, the cells were put into 200 µg/mL hygromycin selection alongside the wild type non-nucleofected cells to assess death kinetics. Selection was continued until the wild type cells were dead (1-2 weeks). Cells were expanded and frozen with plasmid expression assessed by Western blotting.

2.2.6.4 TOPFlash assay

The Cignal TCF/LEF Reporter assay (TOP-GFP) (Qiagen) was used to assess the effect of PADI4 inhibition on Wnt3a driven transcriptional output. Mouse ES cells stably expressing hPADI4 were seeded in six well plates on

day 1. On the following day, 10 µL reporter DNA plasmid was transfected per well using Lipofectamine 2000 as above, except that four hours after transfection with Lipofectamine 2000, cell media was replaced into serum withdrawal media. Cells were treated the morning after transfection, by washing once in PBS and then replacing into Serum withdrawal media prepared with vehicle, 100 ng/mL recombinant Wnt3a (Peprotech), 100 ng/mL recombinant Wnt3a + 200µM Cl-amidine (Cayman Chemical), or 100 ng/mL recombinant Wnt3a + 10µM BB-Cl-amidine (Cayman Chemical) using a P1000 pipette. The negative control FOP plasmid (with mutated and inactive TCF/LEF binding sites, referred to as FOP-GFP) was treated with 100 ng/mL Wnt3a treatment and was additionally included. A GFP-expressing plasmid was used as a positive control to assess approximate transfection efficiency. Three replicate plates were tested which were staggered and treated independently at all stages. After 24 hours treatment, cells were washed in PBS, incubated in 300 µL Accutase, and then inactivated in full serum media. Cells were washed once more in PBS, resuspended in 250µL PBS, and kept on ice for analysis by fluorescence activated cell sorting (FACS) using a BD Accuri C6 flow cytometer. Gating was performed to include live cells. Data analysis gave similar results whether normalised or not against GFP positive cells from the positive control of transfecting GFP included on each plate (~30% of the total population GFP positive and live in single cells after treatment). Representative FACS plots are presented along with average data. Significance was assessed using an unpaired two-tailed t test.

2.2.6.5 Toxicity assay

To test cell toxicity of the cyclic peptides, differentiated HL-60 cells were seeded in a 24 well plate and either DMSO, control peptide_9 or inhibitory peptides 2 and 3 were added. Control peptide_9 was used as a control for the effect of lyophilized cyclic peptides on cell toxicity. After pipetting to mix, 50 µL of cell suspension was mixed in 150 µL Muse assay buffer. The mixture was vortexed just before analyzing using the Muse® Cell Analyzer

with an untreated well of differentiated cells used to gate the machine. Toxicity of peptides was additionally assessed using the Promega CellTox™ Green Cytotoxicity Assay, which was used according to manufacturer's instructions.

2.3 Mass spectrometry methods

2.3.1 General BS^3 conjugation protocol

For optimizing the antibody pulldown, bis(sulfosuccinimidyl)suberate (BS^3) reagent (Thermo Scientific #21580) was used to crosslink the primary antibody to protein A or protein G Dynabeads at the start of the pulldown to minimize the extent of signal from the heavy chain after detection by immunoblot. As BS^3 is moisture-sensitive the vial was let to equilibrate to RT before opening to minimize condensation, with storage at -20˚C. Fresh 5 mM BS^3 solution was prepared in conjugation buffer (250 µL per sample). The Ig-coupled Protein A or Protein G Dynabeads® were washed twice in 400µL BS^3 Conjugation Buffer (20 mM Sodium Phosphate, 0.15M NaCl pH 7.5) by placing on a magnet and discarding the supernatant. The Dynabeads were then incubated in 250 µL 5 mM BS^3 at RT for 30 min with rotation. The cross-linking reaction was quenched in 12.5 µL BS^3 Quenching Buffer (1M Tris-HCl pH 7.5) for 15 min at RT with rotation. The cross-linked Dynabeads were washed three times in 400 µL of IP bind buffer before proceeding with the immunoprecipitation.

2.3.2 Immunoprecipitation and tandem mass spectrometry (IP-MS/MS) protocol for HL60s

25 µL beads per MS/MS condition were washed in buffer (20mM Tris, 150mM NaCl, 0.5% NP-40, with protease inhibitors and PhosSTOP). In 225 µL ND buffer, 3.57 µL PADI4 ab50332 were conjugated per 25 µL washed Dynabeads (ThermoFisher) for 1 hour at RT with rotation. Dynabead washes were performed by placing tubes on a magnet and discarding the supernatant, before resuspending. 5 mM BS^3 solution was prepared freshly in Conjugation Buffer. The immunoglobulin Ig-coupled Dynabeads were then

washed twice in 200 µL Conjugation Buffer and incubated in 250 µL BS^3 (5 mM) for 30 min at RT with rotation. The crosslinking reaction was quenched in 12.5 µL quenching buffer for 15 min at RT with rotation, before washing in 200 µL buffer. Cell lysis began 20 min before the hour-long antibody conjugation was complete. Each T75 flask of 10 $x10^6$ HL60 cells for cell treatment was pelleted, washed in PBS containing protease inhibitors, transferred into Eppendorfs and 100 µL Denaturing buffer was added (20 mM Tris, 1% SDS, 10 mM DTT). Cells were mixed by vortexing for 3 seconds, boiled for 10 min at 95˚C and then diluted into 900 µL buffer. Lysates were passed through a 25 G needle, incubated on ice and 50 µL retained as input in a fresh tube. 950 µL of lysate was added per 50 µL pre-conjugated beads mix (now at 2X volume) and rotated overnight at 4˚C. The supernatant was retained and frozen as flow through along with the input for running by Western blot as required. For submission to Mass spectrometry, beads were then washed three times in ND buffer and twice in 100 mM ammonium bicarbonate (AMBIC) before freezing the beads dry (without supernatant) on dry ice. MS/MS was run at the IGMM Mass spectrometry facility. I then performed subsequent data analysis from the raw data using PEAKS7.5 software.

2.3.3 IP-MS/MS for mES cells stably expressing hPADI4

2.3.3.1 Conjugating antibody or biotinylated peptide 7 to Dynabeads

2.3.3.1.1 Antibody

In low protein-binding tubes, 50 µL Protein A Dynabeads (ThermoFisher) were added per condition, washed in 500 µL PBS + 0.02% Tween-20 and resuspended in 300 µL PBS + 0.02% Tween-20. 1.4 µL hPADI4 ab50332 antibody was added per condition, and incubated for 45 min at RT with rotation before transferring onto ice. Prior to adding cell lysates, beads were washed once in 0.02% Tween-20.

2.3.3.1.2 Biotinylated peptide_7

In low protein-binding tubes, 20 µL Protein A Dynabeads (ThermoFisher) were added per condition, washed in 2 x 300µL PBS + 0.02% Tween-20 and resuspended in 300µL PBS + 0.02% Tween-20. 2.5 µL of biotinylated peptide 7 stock was added per condition, and incubated for 45 min at RT with rotation before transferring onto ice. Beads were then washed twice in PBS, 3 times with 500 µL PBS with 0.1% BSA with a final wash in 500 µL Wash buffer (10 mM Tris-HCl pH 7.5, 150 mM NaCl, with protease inhibitors and PhosSTOP).

2.3.3.2 Cell treatment, lysis, pulldown and washes for interactome analysis

Cells were grown in 10cm^3 dishes and vehicle or 100 µL Cl-amidine (20 mM) (Cayman Chemical) was added to each dish 90 min before cell treatment. Dishes were grown in triplicate and each set of replicates (3 sets of 4 conditions) was staggered by 30 min. Growth media was aspirated and replaced with treatment media: 1) 30 mL full serum media with vehicle; 2) 30 mL full serum media supplemented with 3µM CHIR99021 (AxonMedChem); 3) 30 mL full serum media supplemented with vehicle and 100µL Cl-amidine (20mM) (Cayman Chemical); or 30 mL full serum media supplemented with 3 µM CHIR99021 (AxonMedChem) and 100µL Cl-amidine (20mM) (Cayman Chemical). Cells were incubated for 45 min in treatment media. Treatment medium was then aspirated, and cells washed once in 12 mL ice-cold PBS. 1mL ice cold PBS with protease inhibitors and PhosSTOP was added, cells scraped from dish and transferred to a pre-cooled Eppendorf. Cells were pelleted by centrifugation at 500 x *g* for 3 min at 4˚C and supernatant discarded. Each cell pellet (one per 10cm^3 treated dish) was resuspended by pipetting in 200 µL of ice-cold Lysis buffer (10 mM Tris-HCl pH 7.5, 150 mM NaCl, 0.2% NP-40, with protease inhibitors and PhosSTOP, 25 units/µL Benzonase (Sigma), 2.5mM $MgCl_2$ and 1mM phenylmethylsulfonyl fluoride (PMSF) (Sigma) and incubated on ice for 30 min without rotation, vortexing briefly every 10 min. The cell lysate was then centrifuged at 14,000 rpm for

10min at 4˚C and transfer supernatant to a fresh pre-cooled tube. 400 µL Wash Buffer (10 mM Tris-HCl pH 7.5, 150 mM NaCl (with protease inhibitors and PhosSTOP) was added to each of the tubes of pre-conjugated beads and 190 µL of the cleared lysate and incubated at 4˚C overnight with rotation. Beads were washed three times in Wash buffer and transferred to a new tube. Beads were washed a further two times in 100 mM Ammonium bicarbonate (Sigma), all remaining supernatant was aspirated and beads frozen on dry ice.

2.3.3.3 Cell treatment, lysis, pulldown and washes for PTM analysis

Cell treatments were performed on 12 x10 cm^3 dishes each treated with 10mL full ES serum media supplemented with either 9 µL DMSO vehicle or 9 µL CHIR (from 10 mM stock). Plates were treated in pairs (untreated, treated). Cells were washed once in 10 mL PBS, once in 1mL of PBS supplemented with protease inhibitors and PhosSTOP, scraped and transferred to a fresh Eppendorf. Cells were pelleted by centrifugation at 500 x *g* for 3 min at 4˚C, resuspended in 150 µL Denaturing lysis buffer (20 mM Tris pH 7.5, 1% SDS**,** 10 mM DTT), vortexed for 3 seconds, boiled for 5 min and then kept on ice. The next pair of plates were then harvested. Each condition was diluted in 1350 µL NDh buffer (20 mM Tris pH 7.5, 137 mM NaCl**,** 1% NP-40, with protease inhibitors and PhosSTOP). Benzonase 4 µL and 2 mM $MgCl_2$ was added and samples incubated for 15 min at 4˚C with rotation. Samples were sonicated using the Soniprep150 at amplitude 10 for two cycles of 30 seconds on/off and then incubated for a further 15 min at 4˚C with rotation. Lysates were centrifuged for 5 min at 4400 rpm on bench top centrifuge at 4˚C. 1400 µL lysate was added to a low binding tube containing beads pre-conjugated with antibody or peptide and incubated overnight at 4˚C with rotation. Beads were wash three times in NDh buffer, once in PBS + 0.02% Tween-20 and transferred to a new tube. Beads were then washed twice in 100 mM ammonium bicarbonate, all remaining supernatant was aspirated and beads frozen on dry ice.

2.3.4 Enzyme coverage

Digestion analysis of human PADI4 was performed using the Expasy ProtParam tool from the Swiss Institute for Bioinformatics and using R or Excel, which showed that use of trypsin, chymotrypsin, AspN and Glu-C would improve theoretical coverage from 61.09% using trypsin alone to 97.7% with the four enzymes in combination, assuming 100% digestion efficiency and MS/MS coverage of peptides of between 7-30 amino acids inclusive. An initial multiple enzyme digestion experiment pilot was performed by Dr Clive D'Santos at the CRUK Facility, Cambridge, but digestion efficiencies were found to be low. A modified digestion protocol adapted from *Giansanti et al.* was discussed for future optimization work in-house[4] if and when phosphoproteomic experiments could be established at the same time.

2.3.5 Rapid immunoprecipitation mass spectrometry of endogenous proteins (RIME) methods

Rapid immunoprecipitation mass spectrometry of endogenous proteins (RIME) methods were performed by adaptation from *Mohammed et al*. [5]. Briefly, after cell treatment, each T175 flask of cells was resuspended in 37.5 mL plain Locke's solution and 2.5 mL 16% Methanol free formaldehyde added and incubated for 8 min at RT. 2 x 840 µL of 2.5 M stock glycine was added per condition and mixed to a final 0.1 M concentration in order to quench excess formaldehyde. Cells were pelleted, resuspended in 10 mL PBS supplemented with protease inhibitors into a 15 mL Falcon before washing into 1 mL PBS supplemented with protease inhibitors and continuing with cell lysis and mass spectrometry as used elsewhere in this thesis.

2.3.6 MS/MS procedure

The mass spectrometry experiments were performed by the IGMM Mass spectrometry facility by Dr Niall Quinn or Dr Jimi Wills using the following protocol. Dry frozen beads, prepared as described in Section 2.3.3 above, were incubated for 30 min at 27˚C with trypsin in 2M urea (Sigma), 50 mM Tris pH8 (MRC HGU Technical Services), and the resulting supernatant

transferred to a fresh tube. This was then incubated overnight at 37˚C with 50 µL 50 mM Ammonium bicarbonate, 10% acetonitrile (ACN) (Sigma) and 100 ng modified trypsin. The solution was acidified using 5 µL 10% trifluoroacetic acid (TFA) and diluted to 2% ACN by adding 195 µL 0.1% TFA (Sigma). 100 µL of the resulting peptide solution was loaded onto an activated (using 20 µL methanol), equilibrated (using 100 µL 0.1% TFA) C18 stop-and-go-extraction (STAGE) tip, and washed with 100 µL 0.1% TFA. The bound peptides were eluted into a Protein LoBind 1.5 mL tube (Eppendorf) with 2 x 20 µL 80% ACN 0.1% TFA and concentrated to below 8 µL volume in a vacuum concentrator. The final volume was adjusted to 12 µL in 0.1% TFA. Online liquid chromatography was performed using a Dionex Rapid Separation Liquid Chromatography Nano. Following the C18 clean-up, 5 µL peptides (approximately a fifth of the beads input) were injected twice onto a home-pulled, home-packed C18 analytical column over a gradient of 2%-40% ACN in 40 min, with 0.1% TFA throughout. Eluting peptides were ionised at +2kV before data-dependent analysis on a Thermo Q-Exactive Plus. MS1 spectra were acquired with a mass-to-charge ratio (m/z) range 300-2000 and resolution 70,000, and top 10 ions were selected for fragmentation with normalised collision energy of 30, and an exclusion window of 30 seconds. MS2 spectra were collected with resolution 17,500. The Automatic Gain Control targets for MS1 and MS2 were 3e6 and 1e5 respectively, and all spectra were acquired with lockmass and 1 microscan. An initial analysis was performed by the IGMM facility using MaxQuant and I performed subsequent analysis on the raw data files using both MaxQuant and PEAKS7.5.

2.3.7 MaxQuant analysis

MaxQuant analysis was performed using default settings on raw data generated from the Thermo Q-Exactive Plus with various changes outlined here. Briefly, raw files were loaded and andromeda quantification was used. Search database settings were used from databases previously used at the MRC IGMM MS/MS facility. Multiple enzyme digests were handled using

parameter groups and group-specific parameters, but in general samples were digested with trypsin and this was therefore not required. Label-free quantification (LFQ) was performed with minimum number of neighbours set to 1. The MaxQuant LFQ software performs two searches, firstly to model the error and the second search for quantification using that error. Separate variable modifications were included in both searches. Re-quantification was not used, and matching was performed using 'from and to' settings. MS/MS was not required for LFQ comparison. Contaminants were included so that trypsin levels could be assessed across samples. Default cut-offs were used for identification and match between runs was switched on under the advanced identification settings panel. Fixed modification of carbamidomethylation was typically included as standard in MaxQuant analyses. Variable PTM searches for Citrullinations/ Phosphorylations/ Ubiquitination/ Deamidation(N/Q) result in much longer computational processing time. Peaks7.5 was therefore generally used for PTM analysis and MaxQuant used as an initial way to process samples for analysis across samples by LFQ intensity.

2.3.8 PEAKS analysis

Reference Proteomes for searching fasta files were used from the IGMM MS/MS facility. Identification was performed using precursor mass 8 ppm, fragment mass 0.08 Da, enzyme (trypsin), no non-specific cleavages and allowing 3 missed cleavages. The relevant database was selected using Uniprot format with the following edited rules: Id= ">..\|\(\w+\)" and Desc= ">..\|\w+\|\(.+\)". Quantification was performed with error tolerance 8ppm, RT shift 6 min and default colouring used (red for enrichment, green for depletion using $\log_2$ ratios). Replicates were defined as being in the same group for statistical analysis of differential interacting proteins. Other settings were performed using default settings. Analysis took approximately 24-48 hours. Statistics are reported as $-10\log_{10}(p)$ for which a value >20 is equal to a p-value of <0.01.

2.4 Chemical genetic screen

Various different seeding densities (from 10 up to 10000 cells per well), H3CitR2 antibody concentrations (1:50-1:2000) and fixing methods (ice cold methanol, 4% PFA for 10 min, 4% PFA for one hour) were tested to optimize a high content microscopy acquisition protocol that could detect signal for nuclear H3CitR2 levels in mouse ES cells which grow in 3D colonies in order to set-up a chemical genetic screen using a 96 well plate of inhibitors prepared in DMSO as used in *Williams et al.*[6]. That work had used a high throughput Western method for detection, but an immunofluorescence approach with detection by high content microscopy was taken here. This was adapted from a protocol that was used in *Hari et al*. that had been used to detect the Senescence-Associated Secretory Phenotype from IMR-90 fibroblast cells[7,8]. The resulting optimized protocol for mouse ES colonies is provided. Mock inhibitor plates were prepared by adding DMSO (Hybrid-Max, Sigma) or CHIR99021 (AxonMedChem) with enough to result in a 10μM final concentration in 96 well plates in 1 μL total DMSO and then allowing to dry before freezing the plate. Chemical inhibitor library plates had been provided in the identical format, which was a generous gift from Dr Greg Findlay (MRC PPU) and Prof Nathanael Gray (Harvard University). Mouse ES cells stably expressing human PADI4 were seeded very sparsely at one tenth the usual seeding confluency to a density of 1000 cells per well for treatment the following day. All incubation steps were performed with gentle agitation.

1 μL of DMSO, followed by 100 μL Serum media, was added to each well of the inhibitor plate using a multichannel pipette and shaken for 2 min at RT to dissolve the inhibitor. The 96-well cell plate was aspirated and washed once in PBS. A multichannel Liquidator pipette was then used to transfer the media/inhibitor solution from the inhibitor plate to the cell plate. The 96-well plate was then incubated with inhibitors for 15 min at 5% CO_2 and 37˚C in a cell incubator. Media was then discarded, cells washed twice using the Liquidator pipette with PBS before adding 4% PFA solution in a fumehood with a multichannel pipette to fix the cells for 10 min exactly. After fixation,

the cell plate was washed twice as before in cold PBS. Cells were then permeabilised by incubating in PBS containing 0.2% Tween-20 containing 3% BSA for 1 hour at RT. Cells were then blocked in a PBS solution containing a 0.2% solution of cold water fish skin gelatin (Sigma) and 3% BSA for 1 hour. Cells were then incubated overnight at 4˚C in 1:500 H3CitR2 dissolved in 3% BSA in PBS. In the morning prior to image acquisition, the plate was washed three times in PBS containing 3% BSA and 0.2% Tween-20. Anti-rabbit Alexa 488 was used at 1:1000 dilution in PBS containing 3% BSA and 0.2% Tween-20 prepared immediately before the final plate wash of the cells. 30 µl secondary antibody was added per well before incubation for 30 min in the dark. Cells were then washed three times again in PBS. 30 µl Hoechst 33258 (1 µg/mL in PBS) was then added per well and incubated for 30 min in the dark. Cells were then washed three times again in PBS before acquisition using an Image Xpress Micro XLS. A count of the total number of cells per well is made by counting the nuclei in the 4′,6-diamidino-2-phenylindole (DAPI) channel. The stained area was set to include the nucleus only. Cell parameters were set to determine the minimum and maximum width of identified cells so that positively stained cells could be identified in MetaXpress with appropriate signal intensity above background. The minimum stained area was set to 100 μm^2. Integrated nuclear intensity was normalized to the integrated nuclear intensity for the secondary only antibody control. Alternatively, the average nuclear intensity over two runs was calculated and shown alongside secondary only intensity. An analogous condition at the same seeding density under the same conditions was prepared on a six-well plate and analysed by Western blot as a control for activation. Significance was assessed using an unpaired two-tailed t test.

2.5 Computational methods

2.5.1 Collecting orthologous PADIs

Two methods were used to obtain orthologous PADI sequences from sequence databases. Firstly, a list of proteins was collected with significant similarity (E value < $1x10^{-3}$) to the metazoan PAD_C domain by HMMER

searches against *Reference Proteomes* and *UniProtKB* databases[9]. For completeness, additional more sensitive sequence searches and iterative searches were performed: using tblastn and psiblast against nr/nt; using jackhammer against reference proteomes and *UniProtKB*; and using hhpred against Pfam-A, COG_KOG and PDB_mmCIF70[10-14]. The number of species with a putative PADI orthologues from *UniProtKB* was tabulated along with the total number of species contained in *UniProtKB*. Secondly, the EggNOG database was employed, which, briefly, makes use of a graph-based unsupervised clustering algorithm to infer orthologous groups from 2031 genomes across the tree of life (ENOG410ZKF3: 217 proteins from 74 species) [15]. Phylogenetic reconstruction for the identified PADI orthologues was performed within EggNOG as implemented within the ETE3 suite (eggnog41) and described at http://eggnogdb.embl.de/#/app/methods[15,16].

2.5.2 General phylogenetic tools

For all phylogenetic trees, branch support information was visualized and figures produced using FigTree v1.4.3 and iTOL[17]. Amino acid sequences for PADI homologues were obtained from UniProtKB, NCBI and Pathosystems Resource Integration Center (PATRIC) databases using HMMER and BLAST searches[10,18,19]. PADI2 was used for species with multiple PADI paralogues, as it closest resembles the PADI gene in metazoa with one PADI (such as fish)[20], and with the PADI2 from metazoan species with three PADIs such as birds or reptiles (Figure 3.3).

2.5.2.1 Phylogenetic analysis of other citrullinating enzymes

Sequences of the arginine deiminase from *Giardia lamblia* (gADI)[21] and the porphyromonas-type peptidylarginine deiminase from *Porphyromonas gingivalis* (pPAD)[22] were used as a seed for Hidden Markov Model (HMM) searches of reference proteomes to identify sequences from other species of similar length and most significant similarity[9,18]. These were aligned with 25 representative PADI sequences using MAFFT L-ins-I[23] and singly aligning columns were removed. IQTree was used to produce a maximum likelihood

phylogenetic tree[24,25]. The LG empirical rate matrix with 8 categories of rate variation under the FreeRate model (LG +R8) was used, as determined by ModelFinder[26,27] according to the corrected Akaike Information Criterion. The Ultrafast Bootstrap 2 with 1000 replicates[28], Shimodaira-Hasegawa (SH)-like approximate likelihood-ratio test (aLRT) with 1000 replicates[29-31], and aBayes parametric tests[32] were used to assess node support. The displayed tree is shown unrooted with solid circles indicating consensus node support of >95%.

2.5.2.2 Phylogenetic analysis of PADI orthologues

All putative bacterial PADI sequences in the PATRIC database were obtained from BLAST searches[10,18,19]. In addition, sequences from metazoa were subsampled to maximize the inclusion of different lineages. The human PADI2 sequence was searched against *UniProtKB* to subsample sequences from 35 fungal species that represent the broadest distribution of affinity with PADI2 as determined by HMMER bitscore[9]. This was done to provide representation from the maximum sequence diversity of fungal PADIs. Fragment sequences were excluded. The collected sequences were aligned using MAFFT L-ins-I and singly aligning columns were removed[23]. IQTree was used to produce a maximum likelihood phylogenetic tree[24,25]. The WAG empirical rate matrix with 10 categories of rate variation under the FreeRate model with base frequencies counted from the alignment (WAG +R10 +F) was used as determined by ModelFinder according to the corrected Akaike Information Criterion[27,33]. Ultrafast Bootstrap 2 with 1000 replicates, SH-like aLRT with 1000 replicates, and aBayes parametric tests were used to assess node support[28-32]. The tree is shown rooted at the midpoint with solid circles indicating consensus node support of >95% and a number of critical nodes for testing different evolutionary hypotheses were labelled in full as they are mentioned specifically in the text or referred to in other analyses.

2.5.2.3 Phylogenetic analysis of subsampled PADI orthologues

Bacterial PADI sequences were subsampled to maximize sequence diversity. Both the closest and the most distant bacterial homologues were retained with respect to the metazoan sequence to allow for the broadest distribution of protein sequences (to a total of 50 sequences). Firstly, a maximum likelihood phylogenetic tree was produced using IQTree as above using the best empirical rate matrix according to ModelFinder (WAG + R5 +F0). This reproduced the topology obtained with the full tree. Three further analyses were performed:

1: Maximum likelihood phylogenetic analysis was performed using IQTree under the C20 empirical profile mixture model of evolutionary rate matrices using CIPRES Gateway on XSEDE[34].

2: Bayesian phylogenetic inference was performed using MrBayes v3.2.6 x64 using CIPRES Gateway on XSEDE with mixed model MCMC jumping across different fixed empirical rate matrices and 5 different gamma distributed rate categories[35]. Analysis was performed with 4 runs each of 1000000 chains. The average standard deviation of split frequencies was observed to be <0.005, parameters all had an effective sample size (ESS) > 500 and potential scale reduction factor (PSRF) of 1.000 (to 3 significant figures). The summary tree was generated with a burn-in of 25% over the runs. Posterior probability was used for node support– i.e. where posterior probability was 100, the topology was congruent in every tree sampled by the Markov chain Monte Carlo (MCMC) after burn-in. The *aminoacid model* prior was set as 'mixed' such that the MCMC jumps across different models i.e. mixture of models with fixed rate matrices. Poisson, Jones, Dayhoff, Mtrev, Mtmam, Wag, Rtrev, Cprev, Vt and Blosum models were used and all have equal prior probability. The WAG model had posterior probability of 1.000, and standard deviation <0.0001 – and was exclusively sampled from the posterior[33]. This is consistent with the WAG model being identified as the best empirical matrix identified

according to ModelFinder and the corrected Akaike information criterion from the maximum likelihood analysis in IQTree.

3: Bayesian phylogenetic inference was performed using PhyloBayes under the CAT-GTR model[36,37]. This is an infinite mixture model of rate matrices making use of a Dirichlet process prior. The alignment contains 1100 aligned positions and 50 taxa. 8 chains were performed in parallel for 24 hours such that more than 20000 cycles were achieved as recommended in the PhyloBayes manual using the MRC IGMM and University of Edinburgh computing cluster Eddie3. Readpb, bpcomp, tracecomp tools in PhyloBayes and Tracer software were then used to analyse runs. Posterior consensus trees were generated for each run and were reproducible across the eight different runs. The trace plots for independent runs were also analyzed to assess for apparent stationarity aiming for an ESS of at least 100. Maxdiff was observed to be < 0.1 (maxdiff= 0.06209, meandiff= 0.00330).

Tree topologies were congruent across the four different methods. Ultrafast bootstrap 2 values over 1000 replicates (for the Maximum likelihood methods) or posterior probabilities (for the Bayesian methods) are presented in the table for the nodes labelled in the tree critical to different evolutionary scenarios. Additional topology testing was performed in PAUP*4.0a163 and in IQTree. 100 random trees were generated along with the maximum likelihood constrained tree where fungal and metazoan sequences were constrained to be monophyletic. The SH-test, approximately unbiased (AU) test and expected likelihood weight (ELW) tests were performed and all other alternative trees, including the constraint tree were rejected ($p<0.0001$) [30,38-40].

2.5.2.4 Phylogenetic analysis to exclude synapomorphic regions from the alignment

Phylogenetic analyses from Figure 3.4 were repeated using an alignment with the PAD_N domain removed and with an alignment in which both the PAD_N domain and regions of synapomorphy (Figure 3.6) were removed. Maximum likelihood analysis using IQTree with ModelFinder used to select the best performing fixed empirical rate matrix (WAG + R5 +F0) as above[24,25,27,28]. Topologies were congruent with analysis of the whole alignment and node support values (Ultrafast Bootstrap 2) for the clades labelled in Figure 4 were provided.

2.5.3 Synapomorphy analysis

2.5.3.1 Domain annotation

Pfam annotates three PAD domains in metazoa (PAD_N, PAD_M and PAD_C). Putative locations for the PAD domains within target PADI sequences (from bacteria and fungi) were identified by aligning each target sequence to five metazoan sequences with TCoffee[41]. Putative domain sequence regions were then used as a target query for HMMER or HHPred searches[18,42]. HMMER searches were made against the *UniProtKB* database and HHPred searches were made firstly against a database of HMM profiles of protein domains in the Protein Data Bank (PDB_mmCIF_4_Aug) and secondly against a database of profiles from Pfam (Pfam-A_v31.0)[13]. A PAD_N domain can be found using HHPred in cyanobacterial sequences diverging after *SPM* and *NX* clades but could not be found in fungal, other bacterial species or earlier diverging cyanobacterial PADI sequences. Once individual sequences were identified as possessing a specific domain architecture, multiple sequence alignments were made of groups of sequences with common putative domain architecture and this was used as a query for each type of search.

For the reported E values in Figure 3.5, the following method was used. All sequences from the highlighted clade in the phylogenetic figure were aligned

using TCoffee, PAD_C, PAD_M and PAD_N domains extracted from the cyanobacterial sequences, and secondly PAD_C and PAD_M domains from the clade containing a mixture of bacterial and fungal sequences. This alignment was used as a seed for searches with HHPred against a database of profiles made of the entire human proteome, and against a database of profiles of Pfam domains (Pfam-A_v31.0). HHPred searches were performed using the MPI Bioinformatics Toolkit of the Max Planck Institute for Developmental Biology, Tübingen, Germany[13,43].

2.5.3.2 Multiple sequence alignment of PAD_N domain

Amino acid sequences were aligned using MUSCLE or TCoffee algorithms[41,44] and visualized using Jalview[45]. Putative PAD_N domains from *SPM*/*NX* clade cyanobacterial PADI sequences were identified using HHPred showing significant statistical evidence for affinity (E-value: 2.5E-05)[13,43]. These were aligned with the PAD_N domain from human PADI paralogues and *Rhincodon typus* (whale shark). The alignment was presented with the program Belvu using a colouring scheme indicating the average BLOSUM62 scores (which are correlated with amino acid conservation) of each alignment column: red (>3.5), violet (between 3.5 and 2) and light yellow (between 2 and 0.5)[46]. PsiPred[47] was used to predict secondary structure for the cyanobacterial PAD_N domains (beta sheets). The secondary structure of the PAD_N domain of human PADI2 was identified from the crystal structure (PDB: 4n2a)[48].

2.5.3.3 Synapomorphy of calcium binding sites

Representative fungal, actinobacterial, cyanobacterial, and metazoan PADI sequences were then analysed for the conservation of all of the calcium-binding sites (a minimum of three residues coordinate each calcium binding site) and for other critical residues contained at the active site (Figure 6). PADIs from the following species were used: 1) metazoan PADIs from *Homo sapiens*, *Xenopus laevis*, *Oncorhynchus mykiss*, *Callorhinchus milii*, *Branchiostoma floridae*, *Priapulus caudatus*; 2) cyanobacterial PADIs from

Cyanothece sp. 8801, Stanieria cyanosphaera, Chlorogloeopsis fritschii PCC 6912, Crocosphaera subtropica, Aphanothece sacrum, Cyanothece sp. 7424; 3) fungal PADIs from *Fusarium sp. FOSC 3-a, Periconia macrospinosa, Paracoccidioides lutzii, Blastomyces parvus, Ajellomyces capsulatus, Emmonsia crescens* and; 4) actinobacterial PADIs from *Streptomyces silvensis, Alteromonas lipolytica, Streptomyces sp. 3214.6, Erythrobacter xanthus, Kibdelosporangium aridum, Nocardia brasiliensis ATCC 700358*. Sequences were aligned using MAFFT L-ins-I and compared to functionally annotated regions in *Slade et al*. 2015 and from crystal structures[23,48,49].

2.5.3.4 Structural analyses

Structural homology searches were performed using the Dali server v3.1 with the extracted PAD_C domain used as query[50]. Superposition of known structures was performed in Chimera[51] using the MatchMaker tool[52]. Briefly, the two structures (PDB: 4n2c and 1xkn) were aligned for the best-aligning pair of chains using the Needleman-Wunsch algorithm and BLOSUM62 matrix. A secondary structure score of 30% was included. The superposition was iterated by pruning long atom pairs such that no pair exceeds 2.0 Å.

2.5.4 Divergence time analyses

2.5.4.1 Estimating PADI divergence time by calibrated phylogenetic analysis with metazoan divergence times from the fossil record

BEAST v2.4.8 was used to produce a time tree of the clade of subsampled PADIs from metazoa and the full clade of closest *SPM*/*NX* clade cyanobacteria contained within the PATRIC database using the GTR model with 4 gamma distributed rate categories[53-55]. The following metazoan species were used: *Homo sapiens* (HS), *Mus musculus* (MM), *Alligator mississippiensis* (AM), *Chelonia mydas* (CM), *Gallus gallus* (GG), *Xenopus laevis* (XL), *Oncorhynchus mykiss* (OM), *Callorhinchus milii* (CM), *Branchiostoma floridae* (BF), *Priapulus caudatus* (PC). Node times were set as the following normally distributed priors to calibrate nodes on the tree:

mean 797.0, sigma 72.5 (clade of HS, MM, AM, CM, GG, XL, OM, CM, BF, PC); mean 692.5, sigma 57.5 (clade of HS, MM, AM, CM, GG, XL, OM, CM, BF); mean 473.5, sigma 14.0 (clade of HS, MM, AM, CM, GG, XL, OM, CM); mean 435.0, sigma 6.5 (clade of HS, MM, AM, CM, GG, XL, OM); mean 311.0, sigma 7.5 (clade of HS, MM, AM, CM, GG); mean 89.5, sigma 3.0 (clade of HS, MM). Metazoan divergence times from the fossil record were obtained from timetree.org with bounds on the distributions chosen to span the range of times from the literature centered on the median value[56]. The calibrated Yule model was used as the tree prior. XML files were generated in BEAUti and the MCMC analysis was run using BEAST2 on the CIPRES Gateway on XSEDE. An initial MCMC run of 5,000,000 chains was run for each clock model[57,58]. Then full analysis was run with two independent runs of 10,000,000 chains for the strict clock, uncorrelated lognormal (UCLN) and uncorrelated exponentially distributed (UCED) clock models and three independent runs of 10,000,000 chains for the random local clocks model. All models were additionally run under the tree prior (i.e. in the absence of sequence data) and assessed. Analysis of parameters was performed in Tracer to assess apparent stationarity for the different tree parameters and for acceptable ESS values and congruence was assessed across the independent runs. The predicted divergence time of the metazoan and cyanobacterial clades was given by the marginal posterior distribution of the age of the root of the whole tree. This is given by the TreeHeight parameter. The 95% confidence interval of the TreeHeight parameter for all runs under each of the clock models did not exceed 1300. These data were plotted as a violin plot showing 95% confidence intervals.

2.5.4.2 Accumulated genetic divergence analysis relative to other proteins

Bitscore density is calculated by taking the bitscore of a query sequence to the target sequence produced by hmmer and dividing by the bitscore of the query sequence to itself (longer sequences have higher bitscores), which gives a value between 0 and 1[18]. The bitscore densities of the similarity of 1)

the cyanobacterial homologue to the human sequence: $\Delta bitscoreD_{Cy\text{-}Hu}$ (AC+AH) and 2) of the branchiostomal homologue to the human sequence: $\Delta bitscoreD_{Br\text{-}Hu}$ (XB+XH) were both calculated (Figure 3.9A). A measure of the total accumulated genetic divergence between late-diverging cyanobacteria (*Cyanothece spp.*) and the last common ancestor of *Branchiostoma spp.* and *Homo sapiens* was then calculated by subtracting $\Delta bitscoreD_{Br\text{-}Hu}$ from the $\Delta bitscoreD_{Cy\text{-}Hu}$. This accumulated genetic divergence (AGD) value was calculated for (1) 26 ribosomal proteins (uS2, uS3, uS4, uS5, uS7, uS8, uS9, uS10, uS11, uS12, uS13, uS17, uS19, uL1, uL2, uL3, uL4, uL5, uL6, uL11, uL13, uL14, uL15, uL22, uL23, uL24), (2) 19 sequences whose proteins are mitochondrially located so are reasonable endosymbiont-derived gene transfer candidates (EGT candidates) from the mitochondrion (OTC, ASS1, ARLY, CPS1, PGK, ENO, GAPDH, PK, NAXE, G6PD, RPIA, FUMH, SDHB, SDHA, CS, MDHM, DLAT, DLDH, ACLY), and (3) all 10 proteins still encoded in the mitochondrial genome (MT-ATP6, MT-CO1, MT-CO2, MT-CO3, MT-CYB, MT-1, MT-2, MT-3, MT-4, MT-5)[59]. It is notable that by definition, only very highly conserved proteins have an AGD that can be calculated in this extreme example between the last common ancestor of late diverging cyanobacteria and humans: if a protein has diverged substantially then the similarity of the human homologue to the cyanobacterial will not be discernible and no bitscore can be calculated. AGD values of proteins in each category were then tested for deviation from a normal distribution using the Shapiro Wilk test ($W = b^2 / SS$)[60]. As the calculated p value > 0.05, the null hypothesis was retained and the data treated as being normally distributed. Kurtosis and skew were also within the range of the normal distribution. The AGD for PADI proteins ($AGD_{PADI\ proteins}$ = 0.066) was then compared to the mean AGD of each category of control proteins (e.g. $AGD_{ribosomal\ proteins}$ = 0.70) and z-scores were calculated.

2.5.4.3 Extent of evolution analysis with DNA sequences

Nucleotide sequences for rRNA were obtained from the SILVA database[61]. Nucleotide sequences for PADIs were obtained from NCBI and exons extracted. Comparisons were made with EMBOSS Needle using the Needleman-Wunsch global alignment algorithm (gap open: 10, gap extend: 0.5)[62].

2.6 References for Chapter 2

1. Christophorou, M. A. *et al.* Citrullination regulates pluripotency and histone H1 binding to chromatin. *Nature* **507,** 104–108 (2014).
2. Theunissen, T. W. *et al.* Nanog Overcomes Reprogramming Barriers and Induces Pluripotency in Minimal Conditions. *Current Biology* **21,** 65–71 (2011).
3. Collins, S. J., Ruscetti, F. W., Gallagher, R. E. & Gallo, R. C. Terminal differentiation of human promyelocytic leukemia cells induced by dimethyl sulfoxide and other polar compounds. *PNAS* **75,** 2458–2462 (1978).
4. Giansanti, P., Tsiatsiani, L., Low, T. Y. & Heck, A. J. R. Six alternative proteases for mass spectrometry-based proteomics beyond trypsin. *Nature Protocols* **11,** 993–1006 (2016).
5. Mohammed, H. *et al.* Rapid immunoprecipitation mass spectrometry of endogenous proteins (RIME) for analysis of chromatin complexes. *Nature Protocols* **11,** 316–326 (2016).
6. Williams, C. A. C. *et al.* Erk5 Is a Key Regulator of Naive-Primed Transition and Embryonic Stem Cell Identity. *Cell Rep* **16,** 1820–1828 (2016).
7. Hari, P. & Acosta, J. C. Detecting the Senescence-Associated Secretory Phenotype (SASP) by High Content Microscopy Analysis. *Methods Mol. Biol.* **1534,** 99–109 (2017).
8. Hari, P. *et al.* The innate immune sensor Toll-like receptor 2 controls the senescence-associated secretory phenotype. *Science Advances* **5,** eaaw0254 (2019).
9. Potter, S. C. *et al.* HMMER web server: 2018 update. *Nucl. Acids Res.* **46,** W200–W204 (2018).
10. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* **215,** 403–410 (1990).
11. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25,** 3389–3402 (1997).
12. Söding, J. Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21,** 951–960 (2005).
13. Alva, V., Nam, S.-Z., Söding, J. & Lupas, A. N. The MPI bioinformatics Toolkit as an integrative platform for advanced protein sequence and structure analysis. *Nucl. Acids Res.* **44,** W410–5 (2016).
14. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucl. Acids Res.* **47,** D427–D432 (2019).
15. Huerta-Cepas, J. *et al.* eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucl. Acids Res.* **44,** D286–D293 (2016).
16. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33,** 1635–1638 (2016).
17. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucl. Acids Res.* **44,** W242–W245 (2016).
18. Finn, R. D. *et al.* HMMER web server: 2015 update. *Nucl. Acids Res.* **43,** W30–W38 (2015).
19. Wattam, A. R. *et al.* Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. *Nucl. Acids Res.* **45,** D535–D542 (2017).
20. György, B., Tóth, E., Tarcsa, E., Falus, A. & Buzás, E. I. Citrullination: A posttranslational modification in health and disease. *The International Journal of Biochemistry & Cell Biology* **38,** 1662–1677 (2006).
21. Carolina Touz, M. *et al.* Arginine deiminase has multiple regulatory roles in the biology of Giardia lamblia. *J Cell Sci* **121,** 2930–2938 (2008).
22. McGraw, W. T., Potempa, J., Farley, D. & Travis, J. Purification, Characterization, and Sequence Analysis of a Potential Virulence Factor from Porphyromonas gingivalis, Peptidylarginine Deiminase. *Infect. Immun.* **67,** 3248–3256 (1999).
23. Katoh, K., Rozewicki, J. & Yamada, K. D. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinformatics* **20,** 1160–1166 (2019).

24. Nguyen, L.-T., Schmidt, H. A., Haeseler, von, A. & Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* **32,** 268–274 (2015).
25. Trifinopoulos, J., Nguyen, L.-T., Haeseler, von, A. & Minh, B. Q. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucl. Acids Res.* **44,** W232–W235 (2016).
26. Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25,** 1307–1320 (2008).
27. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Haeseler, von, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* **14,** 587–589 (2017).
28. Hoang, D. T., Chernomor, O., Haeseler, von, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **35,** 518–522 (2018).
29. Shimodaira, H. & Hasegawa, M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* **16,** 1114–1116 (1999).
30. Shimodaira, H. & Hasegawa, M. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* **17,** 1246–1247 (2001).
31. Guindon, S. *et al.* New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Syst Biol* **59,** 307–321 (2010).
32. Anisimova, M., Gil, M., Dufayard, J.-F., Dessimoz, C. & Gascuel, O. Survey of Branch Support Methods Demonstrates Accuracy, Power, and Robustness of Fast Likelihood-based Approximation Schemes. *Syst Biol* **60,** 685–699 (2011).
33. Whelan, S. & Goldman, N. A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. *Mol. Biol. Evol.* **18,** 691–699 (2001).
34. Quang, L. S., Gascuel, O. & Lartillot, N. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* **24,** 2317–2323 (2008).
35. Ronquist, F. *et al.* MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Syst Biol* **61,** 539–542 (2012).
36. Lartillot, N. & Philippe, H. A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process. *Mol. Biol. Evol.* **21,** 1095–1109 (2004).
37. Lartillot, N., Rodrigue, N., Stubbs, D. & Richer, J. PhyloBayes MPI: Phylogenetic Reconstruction with Infinite Mixtures of Profiles in a Parallel Environment. *Syst Biol* **62,** 611–615 (2013).
38. Strimmer, K. & Rambaut, A. Inferring confidence sets of possibly misspecified gene trees. *Proc. Biol. Sci.* **269,** 137–142 (2002).
39. Shimodaira, H. An Approximately Unbiased Test of Phylogenetic Tree Selection. *Syst Biol* **51,** 492–508 (2002).
40. Susko, E. Tests for two trees using likelihood methods. *Mol. Biol. Evol.* **31,** 1029–1039 (2014).
41. Di Tommaso, P. *et al.* T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucl. Acids Res.* **39,** W13–W17 (2011).
42. Söding, J., Biegert, A. & Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucl. Acids Res.* **33,** W244–W248 (2005).
43. Zimmermann, L. *et al.* A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *Journal of Molecular Biology* **430,** 2237–2243 (2018).
44. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl. Acids Res.* **32,** 1792–1797 (2004).
45. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25,** 1189–1191 (2009).
46. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *PNAS* **89,** 10915–10919 (1992).
47. Jones, D. T. Protein secondary structure prediction based on position-specific scoring

matrices. *Journal of Molecular Biology* **292,** 195–202 (1999).
48. Slade, D. J. *et al.* Protein arginine deiminase 2 binds calcium in an ordered fashion: implications for inhibitor design. *ACS Chem. Biol.* **10,** 1043–1053 (2015).
49. Arita, K. *et al.* Structural basis for Ca2+-induced activation of human PAD4. *Nature Structural & Molecular Biology* **11,** 777–783 (2004).
50. Holm, L. & Rosenström, P. Dali server: conservation mapping in 3D. *Nucl. Acids Res.* **38,** W545–9 (2010).
51. Pettersen, E. F. *et al.* UCSF Chimera—A visualization system for exploratory research and analysis. *Journal of Computational Chemistry* **25,** 1605–1612 (2004).
52. Meng, E. C., Pettersen, E. F., Couch, G. S., Huang, C. C. & Ferrin, T. E. Tools for integrated sequence-structure analysis with UCSF Chimera. *BMC Bioinformatics* **7,** 339 (2006).
53. Uyeda, J. C., Harmon, L. J. & Blank, C. E. A Comprehensive Study of Cyanobacterial Morphological and Ecological Evolutionary Dynamics through Deep Geologic Time. *PLoS ONE* **11,** (2016).
54. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7,** 214 (2007).
55. Bouckaert, R. *et al.* BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLOS Computational Biology* **10,** e1003537 (2014).
56. Kumar, S. & Hedges, S. B. TimeTree2: species divergence times on the iPhone. *Bioinformatics* **27,** 2023–2024 (2011).
57. Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. Relaxed Phylogenetics and Dating with Confidence. *Plos Biol* **4,** e88 (2006).
58. Drummond, A. J. & Suchard, M. A. Bayesian random local clocks, or one rate to rule them all. *BMC Biol.* **8,** (2010).
59. Timmis, J. N., Ayliffe, M. A., Huang, C. Y. & Martin, W. F. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat. Rev. Genet.* **5,** 123–135 (2004).
60. Shapiro, S. S. & Wilk, M. B. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika* **52,** 591–& (1965).
61. Yilmaz, P. *et al.* The SILVA and 'All-species Living Tree Project (LTP)' taxonomic frameworks. *Nucl. Acids Res.* **42,** D643–D648 (2014).
62. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* **48,** 443–453 (1970).

Chapter 3: The evolutionary origin of the peptidyl arginine deiminases

3.1.1 Introduction

The evolutionary provenance of the peptidyl arginine deiminase family of enzymes remains unresolved. Paralogues of the peptidyl arginine deiminases appear in bony fish, birds, reptiles, amphibians and mammals, but are missing from many eukaryotes including plants, yeast, and insects. The PADI gene is therefore widely thought to have appeared first in the last common ancestor of teleosteans and mammals[1-4]. Crystal structures hint at a possibly more ancient origin as they reveal that the catalytic domain adopts the same pentein fold as a variety of other widely distributed proteins that otherwise show little similarity in terms of amino acid conservation[5,6]. These proteins include a broad family of guanidino-group (the functional group of the side chain of arginine and agmatine) modifying enzymes with hydrolase, dihydrolase and amidinotransferase catalytic activity. The pentein fold containing enzymes share a broad catalytic core of a Cys, His and two polar guanidine binding residues – Asp or Glu[6]. These residues are arranged within an α/β propeller fold possessing characteristic fivefold rotational pseudosymmetry[5,6].

PADIs comprise three protein domains in total (PAD_N, PAD_M and PAD_C) and bind up to six calcium ions (4 in the PAD_M, and 2 in the PAD_C), which allosterically regulate the catalytic domain (PAD_C) of the enzyme[7,8]. The PAD_M domain consists of a DNA binding p53-like immunoglobulin (Ig) type fold, while the PAD_N domain is a cupredoxin-like fold that has lost metal binding residues. Mammalian genomes encode two of the possible ancient homologues of PADIs' catalytic PAD_C domain (N(G),N(G)-dimethylarginine dimethylaminohydrolase [DDAH] and Glycine amidinotransferase [AGAT]), but their possibly homologous domains are very divergent in sequence identity and catalyse different reactions. DDAH and AGAT also lack the N-terminal and middle domains (PAD_N and PAD_M, respectively). Two other citrullinating enzymes occur in some bacteria and early diverging eukaryotes

(pPAD, gADI). These also possess catalytic domains with a pentein fold, which are distinct from the PAD_C domain. These two enzymes also lack the N-terminal and middle domains (PAD_N, PAD_M), are very sequence divergent, and have different substrate specificity[9,10]. The citrullinating enzyme, pPAD, found in *Porphyrmonas gingivalis,* shows close affinity to the catalytic domain of bacterial and eukaryotic Agmatine deiminases (AgD), while the second, gADI, found in *Giardia lamblia* shows close affinity to free L-arginine deiminases (ADI). Interestingly, both of these citrullinating enzymes are extended variants of AgD and ADI proteins where an additional Ig-like domain is connected to the AgD or ADI core, analogously to PADIs which possess an Ig-like PAD_M domain. However, neither the Ig domains in porphyrmonas-type PAD or *Giardia lamblia* ADI show any affinity to the PAD_M Ig-like domain using profile-to-profile searches[11] or using the Dali server[12], and crystal structures of pPAD do not reveal any structural homology with the Ig-like PAD_M domain found in the PADIs either[13]. The activity of free arginine deiminases or agmatine deiminases (or truncated forms of pPAD or gADI) on protein substrates has not been systematically tested to date so it is not known whether the Ig-like extension is responsible for conferring citrullinating activity. Given that Ig domains are widely distributed, it would not be surprising, however, for an Ig-like extension to confer activity to protein substrates by convergent evolution in tandem with a pentein fold containing catalytic domain as for the vertebrate PADIs[13]. In either case, it is unlikely that these other citrullinating enzymes clarify the evolutionary provenance of the vertebrate peptidyl arginine deiminases.

Lateral or horizontal gene transfer (HGT) is the non-ancestral inheritance of genetic information[14,15]. Outside of the established endosymbiont transfer of mitochondrial genes to the nuclear genome and virus-mediated transfer, evidence as to whether individual xenologous genes with recent or ancient horizontal origin persist in the human genome is contentious[16-21]. In the original draft of the human genome, many xenologous gene candidates were proposed, but this analysis was later rebutted in a string of subsequent

papers[17,22-28]. In 2015, over a decade later, a second list of 145 possible HGT candidates persisting in the human genome was proposed by *Crisp et al.* by a similar method on the basis that a possible homologue could be identified in bacteria but not in multiple *Drosophila* and *Caenorhabditis* species (Additional file 3, Genome Biol. 2015)[18]. This list included the five human *PADI* genes. Although the authors acknowledge that conventional vertical descent could also explain this pattern, they argue that HGT is more parsimonious than the many independent gene losses required. Recently, however, this genome-wide approach to search for possible HGT events in vertebrates was once again disputed (Salzberg, Genome Biol. 2017)[20], and 45 of the highest confidence candidates were reanalyzed and rebutted on a case-by-case basis. In the instance of the *PADI* gene, this reanalysis showed that a *PADI* can also be identified in *Priapulus caudatus* (a marine worm) and therefore that the lack of PADI in at least Drosophila spp. must be explained by gene loss[20]. Salzberg additionally recalculates the HGT index used by *Crisp et al*. for many of the possible HGT candidates, including the PADIs, in light of additional sequences that can be identified. This shows that they no longer pass the original parametric criterion for HGT (which was put forward previously by a subset of the authors in *Crisp et al*.)[18,20,21,29]. While writing, a recent report was also published on the biochemical activity of a possible PAD-like enzyme in some species of fungi[30].

The evolutionary origin of vertebrate PADIs and, therefore, of the introduction of citrullination into the human lineage remain enigmatic. As a result, PADIs merit detailed individual consideration; in particular to identify all bone fide PADI homologues across life, and to consider the proposal that PADIs were horizontally transferred into the vertebrate lineage[18,20]. It was also hoped this might provide insight as to the regulation of the mammalian PADI enzymes in identifying conserved regions in the protein sequence.

3.1.2 Objectives:

- Elucidate the evolutionary origin of the mammalian PADI family.
- Consider in detail the contested possibility of HGT into the vertebrate lineage against vertical descent.
- Test activity and calcium dependence of cyanobacterial PADI.

3.2.1 Identifying orthologous PADIs

I began by collecting all PADI orthologues across life. To do this, all sequences across cellular life in current sequence databases were obtained that contain a PAD_C domain, as defined by having significant sequence similarity using searches with HMMER. This was supplemented by iterative jackhmmer searches and Position-Specific Iterated BLAST (PSI-BLAST) searches as well as tblastn searches of genomic databases. This extended the analysis from Salzberg showing that the earliest diverging animals with a PADI (that have been sequenced to date) are *Branchiostoma belcheri* (a cephalochordate), *Saccoglossus kowalevskii* (a hemichordate) and *Priapulus caudatus* (a marine worm)[20]. Unexpectedly, a large number of related PADI sequences were also identified, not only in bacteria, but also in fungi. This confirms the recent report of biochemical PAD activity in *Aspergillus spp.*[30]. These proteins possess fully conserved catalytic residues, well-conserved calcium-binding residues and well-conserved substrate binding residues annotated from the mammalian PADIs; they are therefore likely to be bona fide PADI enzymes (details in Figure 3.6). To quantify this distribution and assess the number of species with a missing PADI, the table in Figure 3.1A was prepared showing the proportion of species that retain a PADI orthologue out of all sequences contained in *Reference Proteomes* or *UniProtKB*. A second approach using the EggNOG database was also used to collect putative PADIs that it identified as being orthologous, by making use of an unsupervised clustering algorithm of all proteins contained in 2031 genomes across cellular life[31]. Figure 3.1B shows an initial phylogenetic tree of all identified PADI orthologs from the EggNOG database. As described previously[18,20], PADIs are not ubiquitous across the metazoa, instead being

most prevalent across vertebrates and present also in *Priapulus caudatus* (Figure 3.1A). PADIs are also not ubiquitous across bacteria, with cyanobacteria most frequently containing a PADI homologue. None of the known archaea that have been sequenced to date have a detectable PADI homologue.

A

Superdomain	Number of species with a PADI in UniProtKB	Number of species with a proteome in UniProtKB	Percentage of species with a PADI in UniProtKB
Bacteria	211	22311	1.0 %
Cyanobacteria	42	373	11.3 %
Actinobacteria	95	2771	3.4 %
Proteobacteria	53	9440	0.6%
Eukaryotes	293	1374	21.3 %
Animals	154	339	45.4 %
Fungi	139	729	19.1 %
Opisthokonta	293	1068	27.4 %
Diverging before opisthokonta	0	306	0.0 %
Archaea	0	769	0.0 %
Viruses	0	35488	0.0 %

B

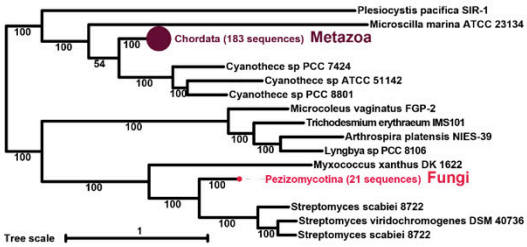


Figure 3.1: Taxon sampling of putative PADI orthologues. **A:** Table showing the number of species possessing a sequence with significant sequence similarity (E value < $1x10^{-3}$) using HMMER searches to the PAD_C domain of vertebrate PADIs from the *UniProtKB* database **B:** Phylogenetic tree of PADI orthologs identified using EggNOG4.5 showing metazoan sequences (183) and fungal sequences (21) each collapsed into a single clade respectively (Section 2.5.1).

An unrooted maximum likelihood phylogenetic tree was then made with an alignment of PADI, ADI, and AgD sequences that included the sequences of the citrullinating enzymes in *Porphyrmonas gingivalis* (pPAD) and *Giardia Lamblia* (gADI) as well as sequences most similar to these citrullinating enzyme variants from HMMER searches. PADIs form a single monophyletic clade: each of the types (PADI, ADI, AgD types) appears to be evolutionarily distinct because they form clear outgroups (Figure 3.2). This is consistent with the annotation in Pfam that pPAD contains an Agmatine deiminase type catalytic domain whereas gADI contains a free arginine deiminase catalytic domain, clearly distinct from the bacterial and fungal PADIs which possess the metazoan PAD_C catalytic domain. This means that a consideration of PADI evolution can be restricted to the previously described animal sequences and the uncharacterized fungal and bacterial PADI sequences.

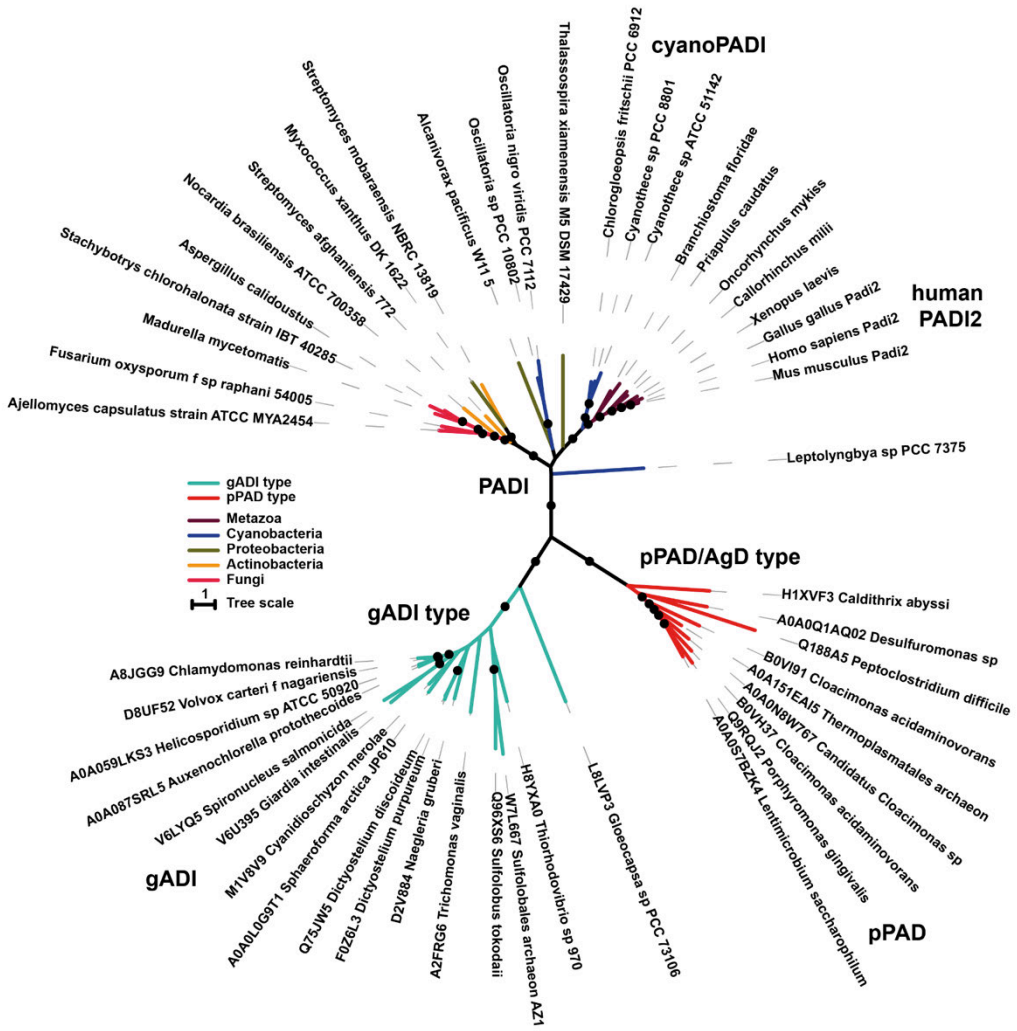


Figure 3.2: PADIs are distinct from ADIs, AgDs, gADI (arginine deiminase from *Giardia lamblia*) and pPAD (porphyromonas-type peptidylarginine deiminase from *Porphyromonas gingivalis*) sequences. The pPAD sequence from *Porphyromonas gingivalis* was searched against *UniProtKB* using HMMER to obtain 8 sequences with highest affinity. The gADI sequence from *Giardia lamblia* was similarly searched against *UniProtKB* using HMMER.15 sequences with high levels of similarity were obtained. These were aligned with 25 PADI sequences subsampled from different domains of life using MAFFT L-ins-I and singly aligning columns were removed. IQTree was used to produce a maximum likelihood phylogenetic tree using the LG empirical rate matrix with 8 categories of rate variation under the FreeRate model (LG +R8) as determined by ModelFinder. Ultrafast Bootstrap 2 with 1000 replicates, SH-like aLRT with 1000 replicates and aBayes parametric tests were used to assess node support (). The tree is shown unrooted with solid circles indicating consensus node support of >95%. Full methods are in Section 2.5.2.1.

This phyletic distribution is consistent with an evolutionary model in which PADI genes are products of vertical evolution but were lost independently in many separate lineages[20]. In this scenario, gene loss would need to have occurred in all early-branching lineages leading to at least 306 non-opisthokont eukaryotes and in other lineages, for example those leading to Drosophila and Caenorhabditis.

3.2.2 Phylogenetic analysis of PADI orthologues

To resolve patterns of gene loss and gain across the tree of life, detailed phylogenetic analysis of PADI sequences was undertaken. Firstly, a large number of PADI orthologue sequences were aligned and a maximum likelihood approach was used to produce an initial phylogenetic tree. Very strong bootstrap support (>95%) was obtained for a clade restricted to cyanobacterial and animal PADIs, which excludes a fully supported outgroup clade containing fungal, actinobacterial and proteobacterial sequences (Figure 3.3). Full node support placed fungal PADIs in a clade with actinobacterial sequences. This tree topology was surprising as it is inconsistent with the known species tree: animal sequences have closer affinity to those in cyanobacteria than they have to fungal sequences.

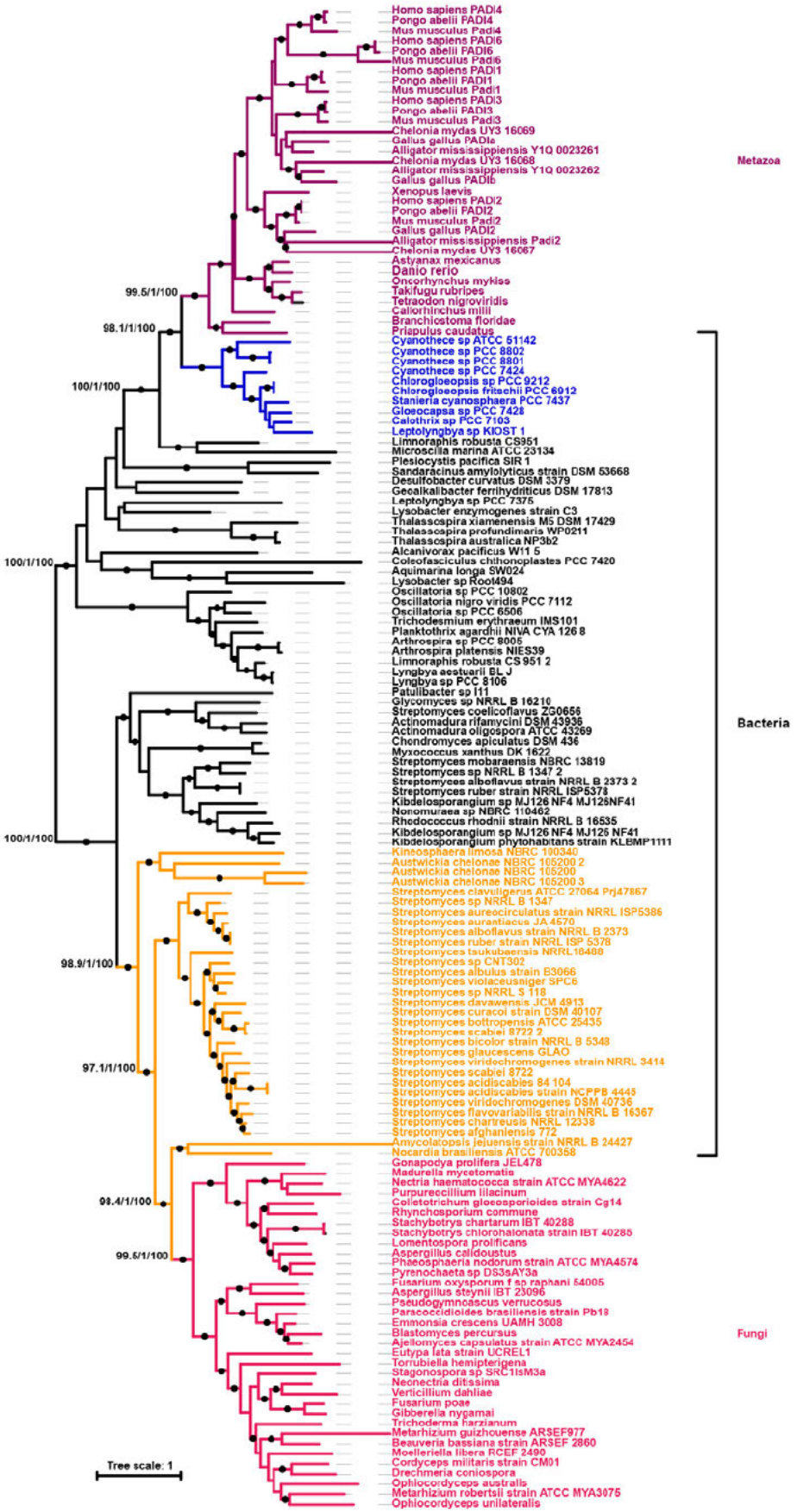
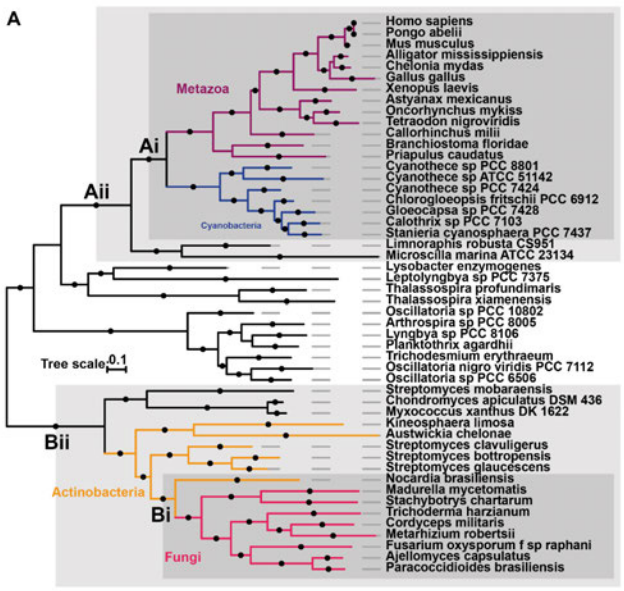

Homo sapiens PADI4
Pongo abelii PADI4
Mus musculus Padi4
Homo sapiens PADI6
Pongo abelii PADI6
Mus musculus Padi6
Homo sapiens PADI1
Pongo abelii PADI1
Mus musculus Padi1
Homo sapiens PADI3
Pongo abelii PADI3
Mus musculus Padi3
Chelonia mydas UY3 16069
Gallus gallus PADI4
Alligator mississippiensis Y1Q 0023261
Chelonia mydas UY3 16068
Alligator mississippiensis Y1Q 0023262
Gallus gallus PADI6
Xenopus laevis
Homo sapiens PADI2
Pongo abelii PADI2
Mus musculus Padi2
Gallus gallus PADI2
Alligator mississippiensis Padi2
Chelonia mydas UY3 16067
Astyanax mexicanus
Danio rerio
Oncorhynchus mykiss
Takifugu rubripes
Tetraodon nigroviridis
Callorhinchus milii
Branchiostoma floridae
Priapulus caudatus
Cyanothece sp ATCC 51142
Cyanothece sp PCC 8802
Cyanothece sp PCC 8801
Cyanothece sp PCC 7424
Chlorogloeopsis sp PCC 9212
Chlorogloeopsis fritschii PCC 6912
Stanieria cyanosphaera PCC 7437
Gloeocapsa sp PCC 7428
Calothrix sp PCC 7103
Leptolyngbya sp KIOST 1
Limnoraphis robusta CS951
Microcoleus marina ATCC 23134
Plesiocystis pacifica SIR 1
Sandaracinus amylolyticus strain DSM 53668
Desulfobacter curvatus DSM 3379
Geoalkalibacter ferrihydriticus DSM 17813
Leptolyngbya sp PCC 7375
Lysobacter enzymogenes strain C3
Thalassospira xiamenensis M5 DSM 17429
Thalassospira profundimaria WP0211
Thalassospira australica NP3b2
Alcanivorax pacificus W11 5
Coleofasciculus chthonoplastes PCC 7420
Aquimarina longa SW024
Lysobacter sp Root494
Oscillatoria sp PCC 10802
Oscillatoria nigro viridis PCC 7112
Oscillatoria sp PCC 6506
Trichodesmium erythraeum IMS101
Planktothrix agardhii NIVA CYA 126 8
Arthrospira sp PCC 8005
Arthrospira platensis NIES39
Limnoraphis robusta CS 951 2
Lyngbya aestuarii BL J
Lyngbya sp PCC 8106
Patulibacter sp I11
Glycomyces sp NRRL B 16210
Streptomyces coelicoflavus ZG0656
Actinomadura rifamycini DSM 43936
Actinomadura oligospora ATCC 43269
Chondromyces apiculatus DSM 436
Myxococcus xanthus DK 1622
Streptomyces mobaraensis NBRC 13819
Streptomyces sp NRRL B 1347 2
Streptomyces alboflavus strain NRRL B 2373 2
Streptomyces ruber strain NRRL ISP5378
Kibdelosporangium sp MJ126 NF4 MJ126NF41
Nonomuraea sp NBRC 110462
Rhodococcus rhodnii strain NRRL B 16535
Kibdelosporangium sp MJ126 NF4 MJ126 NF41
Kibdelosporangium phytohabitans strain KLBMP1111
Kineosphaera limosa NBRC 100340
Austwickia chelonae NBRC 105200 2
Austwickia chelonae NBRC 105200
Austwickia chelonae NBRC 105200 3
Streptomyces clavuligerus ATCC 27064 Prj47867
Streptomyces sp NRRL B 1347
Streptomyces aureocirculatus strain NRRL ISP5386
Streptomyces aurantiacus JA 4570
Streptomyces alboflavus strain NRRL B 2373
Streptomyces ruber strain NRRL ISP 5378
Streptomyces tsukubaensis NRRL18488
Streptomyces sp CNT302
Streptomyces albulus strain B3066
Streptomyces violaceusniger SPC6
Streptomyces sp NRRL S 118
Streptomyces davawensis JCM 4913
Streptomyces curacoi strain DSM 40107
Streptomyces bottropensis ATCC 25435
Streptomyces scabiei 8722 2
Streptomyces bicolor strain NRRL B 5348
Streptomyces glaucescens GLAO
Streptomyces viridochromogenes strain NRRL 3414
Streptomyces scabiei 8722
Streptomyces acidiscabies 84 104
Streptomyces acidiscabies strain NCPPB 4445
Streptomyces viridochromogenes DSM 40736
Streptomyces flavovariabilis strain NRRL B 16367
Streptomyces chartreusis NRRL 12338
Streptomyces afghaniensis 772
Amycolatopsis jejuensis strain NRRL B 24427
Nocardia brasiliensis ATCC 700358
Gonapodya prolifera JEL478
Madurella mycetomatis
Nectria haematococca strain ATCC MYA4622
Purpureocillium lilacinum
Colletotrichum gloeosporioides strain Cg14
Rhynchosporium commune
Stachybotrys chartarum IBT 40288
Stachybotrys chlorohalonata strain IBT 40285
Lomentospora prolificans
Aspergillus calidoustus
Phaeosphaeria nodorum strain ATCC MYA4574
Pyrenochaeta sp DS3sAY3a
Fusarium oxysporum f sp raphani 54005
Aspergillus steynii IBT 23096
Pseudogymnoascus verrucosus
Paracoccidioides brasiliensis strain Pb18
Emmonsia crescens UAMH 3008
Blastomyces percursus
Ajellomyces capsulatus strain ATCC MYA2454
Eutypa lata strain UCREL1
Torrubiella hemipterigena
Stagonospora sp SRC1lsM3a
Neonectria ditissima
Verticillium dahliae
Fusarium poae
Gibberella nygamai
Trichoderma harzianum
Metarhizium guizhouense ARSEF977
Beauveria bassiana strain ARSEF 2860
Moelleriella libera RCEF 2490
Cordyceps militaris strain CM01
Drechmeria coniospora
Ophiocordyceps australis
Metarhizium robertsii strain ATCC MYA3075
Ophiocordyceps unilateralis
99.5/1/100
98.1/1/100
100/1/100
100/1/100
100/1/100
98.9/1/100
97.1/1/100
98.4/1/100
99.5/1/100
Tree scale: 1
Metazoa
Bacteria
Fungi

Figure 3.3: Phylogenetic analysis of putative PADIs. All putative bacterial PADI sequences in the PATRIC database were obtained. In addition, sequences from metazoa were subsampled to maximize the inclusion of different lineages. The human PADI2 sequence was searched against *UniProtKB* to subsample 35 fungal sequences representing the broadest distribution of HMMER affinity. Sequences were then aligned using MAFFT L-ins-I and singly aligning columns were removed. IQTree was used to produce a maximum likelihood phylogenetic tree using the WAG empirical rate matrix with 10 categories of rate variation under the FreeRate model with base frequencies counted from the alignment (WAG +R10 +F) as determined by ModelFinder. Ultrafast Bootstrap 2 with 1000 replicates, SH-like aLRT with 1000 replicates and aBayes parametric tests were used to assess node support. The tree is shown rooted at the midpoint with solid circles indicating consensus node support of >95% and a number of critical nodes labelled in full. Full methods are in Section 2.5.2.2.

Phylogenetic methods, in particular single gene phylogenetic trees, are not infallible. As such the topology may simply represent the failure of phylogenetic inference in the case of this individual gene, such that an artifact, such as model misspecification, might explain the affinity of the separate eukaryotic PADIs to different bacterial PADI types. This might be indicated for example by low bootstrap values or a changing tree topology under different models of rate variation. A common approach for an initial maximum likelihood analysis uses a best-fitting fixed rate matrix of amino acid substitution rates (derived empirically from large datasets of proteins) to produce the tree[32-34]. This could be potentially confounded if there is evolutionary rate variation over different parts of the tree or deviation from typical protein substitution rates. In particular, attention has been drawn in the literature to heterotachous evolution where the evolutionary substitution rate of a given site may change over time, and also to covarion evolution where rate changes over time in one site may be dependent on the rate changes at other sites in a protein[35]. Long-branch attraction, the most famous confounding effect on phylogenetic tree topologies, is a special example of heterotachy in which fast evolving branches that may be unrelated erroneously branch together in a tree topology[36]. The corollary effect of long-branch repels can also be demonstrated due to covarion

evolution[37]. Heterotachy might be particularly plausible in the PADI example given the very divergent species with orthologues under analysis (animal, fungal, cyanobacterial).

To validate the tree topology, and informed by the large tree, PADI sequences with decreasing HMMER bitscore similarity to a profile from metazoan PADIs were therefore subsampled to cover a broad representation of homologue sequences for more computationally expensive phylogenetic analyses. Three approaches were used to address possible heterotachous effects: a Bayesian approach with MCMC model jumping that enables the MCMC process to sample over various different fixed empirical rate matrices[38], a maximum likelihood approach using a mixture model of 20 different fixed amino acid rate matrices (C20)[39], and a Bayesian approach that allows for infinite mixture model categories by making use of a Dirichlet process prior (CAT-Poisson and CAT-GTR)[40]. These methods have been shown to be more robust to long-branch artifacts and more broadly to saturated sequence artifacts due to the more sophisticated treatment of across-site rate heterogeneity[36,41-43].



Method:	Node support for clade:			
	Ai	Aii	Bi	Bii
MrBayes Mixed model jump	100	100	100	100
IQTree WAG+R5+F0	96	100	100	100
IQTree C20 Mixture model	99	100	99	100
PhyloBayes CAT-GTR	99	100	100	100

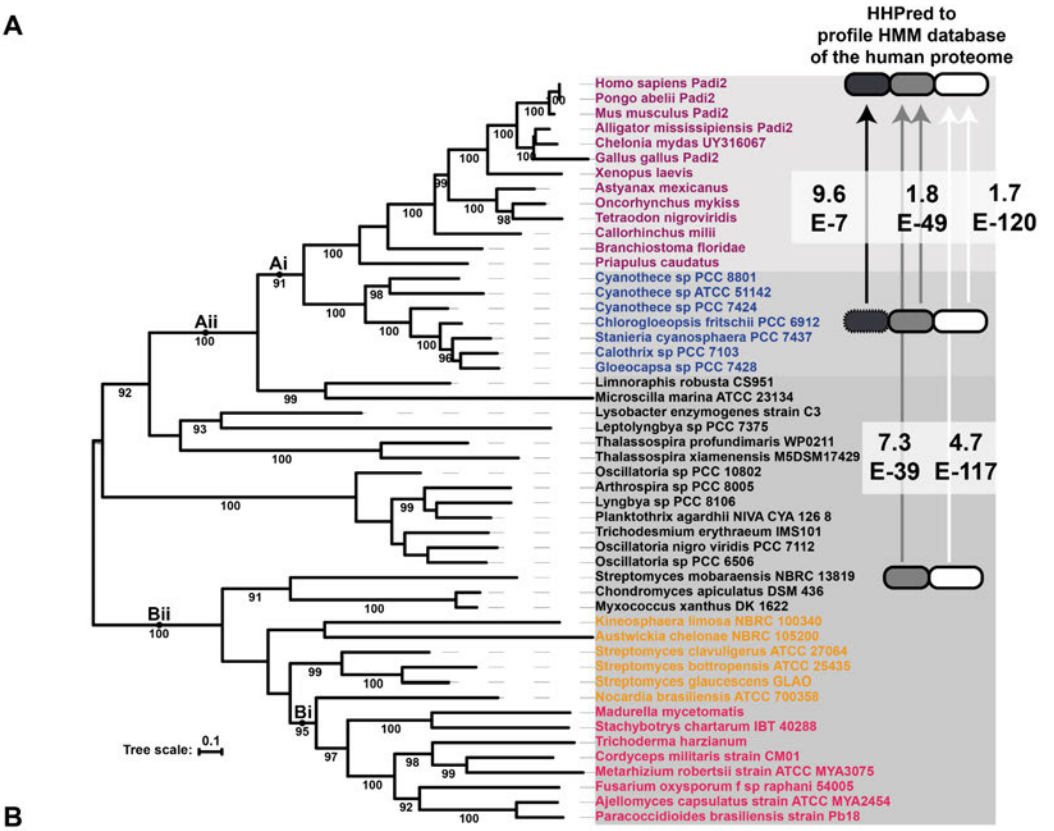
Figure 3.4: Subsampled phylogenetic analysis with different evolutionary rate models. **A:** Bayesian phylogenetic inference using MrBayes with mixed model jumping across different fixed empirical rate matrices (predominantly sampling across the WAG model); **B:** Maximum likelihood inference using IQTree with ModelFinder with the best performing fixed

empirical rate matrix (WAG + R5 +F0); **C:** Maximum likelihood inference using IQTree with the C20 mixture model of rate matrices; and **D:** a Bayesian phylogenetic inference using the CAT-GTR model with an infinite mixture model of rate matrices by using a Dirichlet process prior. Tree topologies were congruent across the different rate matrix models. Ultrafast boostrap 2 values over 1000 replicates (for the Maximum likelihood methods) or posterior probabilities (for the Bayesian methods) are presented in the table for the nodes labelled in the tree. Full methods are in Section 2.5.2.3.

In all of the analyses, only a single topology was recovered that supports a clade of cyanobacterial and animal sequences to the exclusion of a clade of fungal and actinobacterial sequences (Figure 3.4). Posterior probabilities or bootstrap values for this topology were high, approaching 100% for each of the four diverse methods (Ai, Aii: Figure 3.4). The analysis was also repeated using other bootstrap algorithms including the full non-parametric bootstrap, again obtaining full support[44,45]. Topology constraint tests rejected a number of randomly generated trees to rule out possible specific biases in bootstrap resampling. Lastly, a constrained tree for the expected model of vertical evolution was generated where eukaryote PADIs were restricted to a monophyletic group. As expected from the degree of node support in the original trees, these alternative trees and constraint tree were all significantly rejected ($p<0.0001$) by multiple statistical tests (Section 2.5.2.3)[46-48].

3.2.3 Protein domain analysis of PADI orthologues

An analysis was then undertaken to assess how the PADI protein domain architecture is distributed across orthologues. This was firstly done using Pfam annotations, which are powered by HMMER searches[49] and can be found online. All metazoan PADIs possess three Pfam domains, annotated in Pfam as PAD_N, PAD_M and PAD_C domains. The closest cyanobacterial PADIs appear to possess two Pfam domains, a mammalian PAD_M domain and a PAD_C domain, but do not possess the PAD_N domain. More distant cyanobacterial PADI sequences and other bacterial and fungal PADIs, by contrast, are not annotated with either the PAD_N or PAD_M domain.



B

Figure:	Cartoon:	Region of alignment used for phylogenetic analysis:	Node support for clade:			
			Ai	Aii	Bi	Bii
Fig 4		Whole protein	99	100	99	100
not shown		PAD_N domain removed	97	100	99	100
above	* *	PAD_N domain and synapomorphic regions removed	91	100	95	100

C

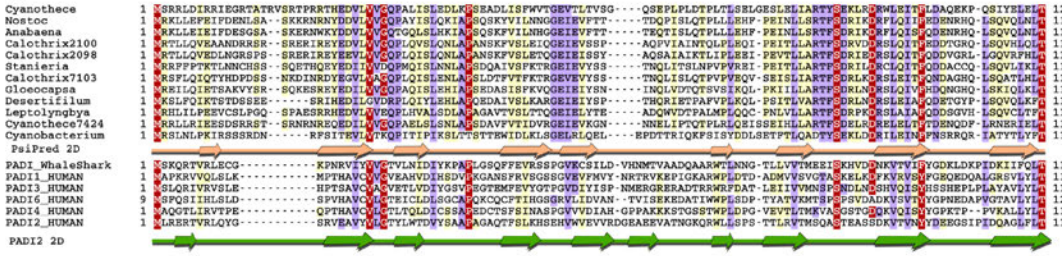


Figure 3.5: Domain architecture analysis of PADI orthologues. **A and B:** Phylogenetic analyses from Figure 3.4 were repeated using an alignment with the PAD_N domain removed (**middle row**), and with an alignment in which both the PAD_N domain and regions of synapomorphy were removed (**A, bottom row**). Maximum likelihood inference using IQTree was used in all three cases and ModelFinder was used to select the best performing fixed empirical rate matrix (WAG + R5 +F0). Topologies were congruent with analysis of the whole alignment (**top row**) so the node support values (Ultrafast Bootstrap 2) for the clades labelled in Figure 3.4 have been provided in the table for the trees produced using the truncated alignments. **A (right hand panel):** The region aligning to the metazoan domains was extracted from sequences belonging to the clades (firstly of cyanobacterial sequences,

and secondly of a mixture of bacterial and fungal sequences) that had been used for phylogenetic analysis. An HMM profile was made and searched with HHPred against a database of profiles made of the entire human proteome, and against a database of profiles of Pfam domains. The E values are given for these searches where significant sequence similarity could be identified from HHPred searches. **C:** Putative PAD_N domains from *SPM/NX* clade cyanobacterial PADI sequences were identified using HHPred showing significant statistical evidence for affinity (**A**). These were aligned with the PAD_N domain from human PADI paralogues and *Rhincodon typus* (whale shark). The alignment was presented with the program Belvu using a colouring scheme indicating the average BLOSUM62 scores (which are correlated with amino acid conservation) of each alignment column: red (>3.5), violet (between 3.5 and 2) and light yellow (between 2 and 0.5). Peach arrows shown below the cyanobacterial sequences indicate PsiPred predicted secondary structure (beta sheets). Green arrows (beta sheets) correspond to the known secondary structure of the PAD_N domain of human PADI2 as identified from the crystal structure (PDB: 4n2a). Figure in 3.5C was prepared with Dr Luis Sanchez-Pulido, who also assisted with HHPred analysis.

Manual inspection revealed possibly significant alignment in the N-terminus of the bacterial and fungal sequences with animal PADIs. Consequently, more sensitive profile-to-profile HMM searches[11] were used to explore this and to identify domains that might have been overlooked by Pfam. A multiple sequence alignment was made firstly of cyanobacterial species contained in the monophyletic clade of metazoan sequences (Figure 3.5A, sequences coloured in dark blue), and secondly of the remaining bacterial and fungal sequences (Figure 3.5A, sequences coloured in black, orange and pink). By comparing the alignment with crystal structures of human PADI2 and PADI4, regions corresponding to each of the three Pfam domains were extracted and used as a seed for HHPred searches against a database of profiles of the human proteome and a database of profiles of all domains contained in Pfam. These results revealed that the outgroup bacterial and fungal sequences additionally possess a more divergent version of the PAD_M domain, but no PAD_N domain – the PAD_N region is completely absent in those bacterial and fungal orthologues. The closest cyanobacterial homologues, by contrast, possess both a PAD_M domain (with much closer affinity than the outgroup sequences), as well as a degenerate metazoan

PAD_N cupredoxin type domain. Although not identifiable by HMMER, this was identified as significantly sequence similar using HHPred (E value<1x10^{-7}) (Figure 3.5A, right panel). With Dr Luis Sanchez-Pulido, PsiPred was then used to predict the secondary structure of the PAD_N domain from three-domain cyanobacterial sequences[50]. With Dr Luis Sanchez-Pulido, a multiple sequence alignment of the cyanobacterial PAD_N region with the metazoan PAD_N was prepared using Belvu[51] (Figure 3.5C). This shows that the secondary structure of the cyanobacterial PAD_N domain predicted using PsiPred aligns well with the experimentally determined secondary structure of the human PAD_N domain from PADI2 derived from X ray crystal structure data (Figure 3.5C).

A synapomorphy is a character or sequence feature, different from that of an ancestor, which distinguishes members of a monophyletic clade from other organisms in different clades. In this instance, late-diverging cyanobacterial and animal PADIs share a synapomorphic protein domain (PAD_N) that is absent from earlier diverging cyanobacterial, other bacterial and fungal sequences. Taken on its own, this is a very surprising result as it indicates cyanobacterial and metazoan PADIs comprise a monophyletic group to the exclusion of fungal and bacterial sequences. This evidence corroborates the topology derived from phylogenetic analysis. Phylogenetic analyses were then repeated on a multiple sequence alignment where the PAD_N domain was removed. The same topology was recovered with high bootstrap values showing that the two lines of evidence (phylogenetic and PAD_N domain analysis) are independent from each other (Figure 3.5B).

3.2.4 PADI orthologue protein features and synapomorphy

In light of the above findings using the PAD_N domain, a similar analysis was undertaken to analyse calcium-binding residues and other annotated functional features of the mammalian PADI sequence for their conservation and distribution in bacterial and fungal orthologues. PADIs require very high calcium concentrations for catalytic activity *in vitro*, and even single point

mutants of calcium binding residues can abolish activity[7,8]. At high calcium concentrations the catalytic nucleophilic cysteine residue moves as much as 5-12 Å into its catalytically competent conformation. The allosteric binding of up to six calcium ions (five in PADI4) thereby allows formation of the active site cleft and is required for catalytic activity. In detailed analysis of PADI2, crystal structures were solved at eight different calcium concentrations[8]. This, together with important earlier work on PADI4[7], revealed an exquisite mode of allosteric regulation by sequential calcium ion binding. At low calcium concentrations, high affinity calcium site 6 in the PAD_M domain (specific to PADI2, not conserved in PADI4) and 1 in the PAD_C domain retain calcium binding, but all other calcium sites remain uncoordinated. Operating as a "calcium switch", binding sites 3-5, located within residues 153-179 in the PAD_M domain, show sequential binding of calcium (3 and 5, then 4) at increasing calcium concentrations. Of these, calcium site 4 derives one conserved binding residue from the PAD_C domain (D389), distant from its other coordinating residues in linear sequence. This is used to explain how the binding sites in the PAD_M domain can communicate allosterically with the distant PAD_C domain. Finally, calcium-binding site 2, also located in the PAD_C domain and adjacent in linear sequence to a loop that contains D389 from Ca site 4, becomes occupied only at the highest calcium concentrations (10 mM) in the fully active holo-conformation (fully calcium occupied structure) of human PADI2. A similar full occupancy was observed in the earlier crystal structure of calcium saturated human PADI4. The final calcium binding (Ca2), in particular, stabilises the catalytically competent conformation, which results in a fully structured active site cleft. In the catalytic conformation, the nucleophilic cysteine moves ~5 Å as compared to the low calcium enzyme conformation. Representative fungal, actinobacterial, cyanobacterial, and metazoan PADI sequences were then analysed for the conservation of all of the calcium-binding sites (a minimum of three residues coordinate each calcium binding site) and for other critical residues contained at the active site (Figure 3.6).

All catalytic residues and substrate binding residues are fully conserved among all PADI homologues (Figure 3.6). In addition, calcium-binding sites 3 and 1 appear to be fully conserved, while calcium site 5 is likely to be mostly conserved as well. Functionally, Ca6 is likely to be conserved as the substitution of D125 to N and E131 to D, which is present in both actinobacterial and fungal sequences, is likely to preserve ion binding. Intriguingly, however, calcium sites 2 and 4 appear to be conserved only in cyanobacterial and metazoan sequences. The fungal and actinobacterial sequences diverge from binding sites 2 and 4 to a different amino acid motif in common between these other bacterial and fungal homologues, which may or may not retain the calcium coordinating function. Critically, only cyanobacterial and animal sequences conserve the calcium switch residue D389 (residues: 369-389) (substituted to Gly in both actinobacterial and fungal sequences, and therefore chemically inequivalent and incompetent for metal coordination). This indicates that the ordered, sequential calcium binding in the PAD_M domain that is responsible for the allosteric communication with the catalytic PAD_C domain in human PADI2[8] is likely only to be conserved in cyanobacterial and metazoan PADIs. As a result, a potentially different mode of calcium regulation operates in the fungal and actinobacterial PADIs from that present in cyanobacterial and metazoan PADIs.

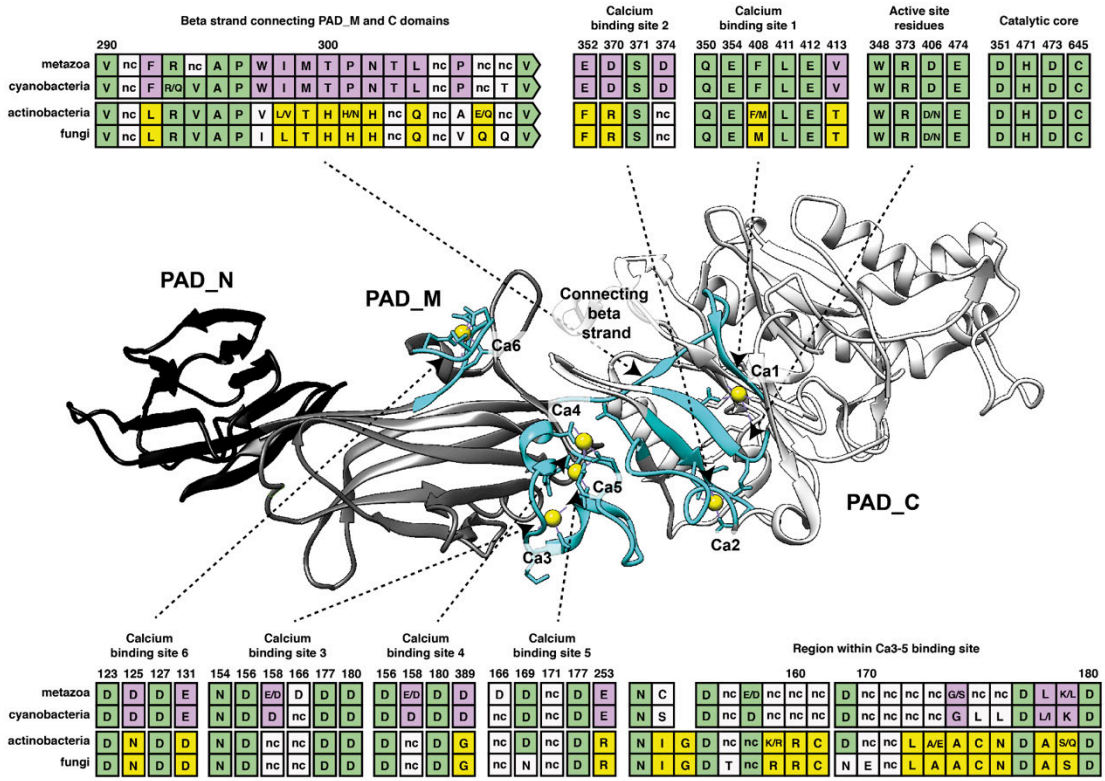


Figure 3.6: Analysis of synapomorphic regions. Six PADI sequences from each of metazoa, cyanobacteria, actinobacteria and fungi were aligned using MAFFT L-ins-I. Consensus sites across the six species or sites shared across two amino acids are shown with standard single letter amino acid abbreviations. In the absence of consensus conservation to one or two amino acids across the species, the site is shown as "nc". Numbering is given above the alignment and corresponds to the ungapped site of human PADI2 so that residues can be compared to *Slade et al.*[8]. Sites showing conservation across all four domains are coloured in green, sites showing synapomorphy between metazoa and cyanobacteria are coloured in purple, and sites showing synapomorphy between fungi and actinobacteria are coloured in yellow. A figure of the crystal structure of PADI2 from Homo sapiens is also presented with PAD_N domain coloured in black, PAD_M domain in grey and PAD_C domain in white. Synapomorphic regions are coloured in cyan and calcium ions are shown as yellow spheres. Methods are in Section 2.5.3.3.

Looking in close detail at regions specific to the outgroup sequences revealed that fungal and actinobacterial sequences additionally share features that are not present in the metazoan and cyanobacterial PADIs. This includes a conserved region within calcium binding site 3-5 that is absent from the metazoan and cyanobacterial sequences (Figure 3.6: amino acids 155-180, where differences conserved between fungal and actinobacterial

sequences are highlighted in yellow). Also of interest is a highly conserved 10 amino acid beta sheet that connects the PAD_M and PAD_C domains and which is conserved in full in the cyanobacterial and metazoan sequences (Figure 3.6: amino acids 292-302). In fungal and actinobacterial sequences the region is also conserved closely, but to a different 10 amino acid sequence containing a distinctive triple histidine motif (Figure 3.6, amino acids 300-302).

This analysis therefore reveals features shared between cyanobacterial and animal sequences that differ or are absent from fungal and bacterial outgroups. It additionally reveals features shared only between fungal and actinobacterial sequences and missing from metazoan and cyanobacterial PADIs. The topology of phylogenies built either with or without these synapomorphic sequence features was congruent (Figure 3.5B), indicating that the synapomorphies represent independent additional support for the closer affinity of vertebrate and cyanobacterial PADIs than to their other homologues, just as for the analysis using the PAD_N domain. As these occur at the level of the amino acid sequence and at the level of a protein domain, they are robust to convergent evolution[52] and to differences in rate variation across the tree, and therefore provide strong support of the phylogenetic topology (Figure 3.3). It is implausible that these blocks of sequence of up to ten amino acids were derived by chance, independently, in actinobacterial and fungal PADIs and thereby act as synopomorphic motifs as they imply common ancestry to actinobacterial and fungal PADIs that is distinct from the ancestry of cyanobacterial and metazoan PADIs.

3.2.5 Molecular clock analysis and divergence time analysis

Evidence presented thus far can be reconciled with evolutionary vertical descent if the last universal common ancestor (LUCA), which lived at least 3.35 billion years ago[53], harboured two paralogous PADI genes which then, as they were transmitted to bacterial, archaeal and eukaryotic lineages, were largely deleted except for vertebrate, fungal, actinobacterial and

cyanobacterial lineages in which one, but never both, were retained. In this hypothetical scenario, there are two large gaps where PADI paralogues are missing from modern genomes and must have been independently lost twice. Firstly, a missing PADI of the three-domain cyanobacterial/metazoan type was lost from lineages leading to every other species in life. Secondly a missing two-domain PADI of the fungal/actinobacterial type must be separately accounted for in independent gene losses in lineages leading to all other species except for those fungi and bacteria that maintain it (Figure 3.1A).

Evidence for this highly unparsimonious model of vertical descent would be rates of PADI sequence evolution, bridging the gaps in the species phylogeny where PADI orthologues are missing, that do not appear to be anachronistic either with respect to (1) geologically defined timings, or (2) other genes that are known to have been acquired vertically from the LUCA. By contrast, strong additional evidence for horizontal gene transfer would be anomalously slow rates of PADI sequence evolution across species bridging the hypothesized gap. It is worth noting that PADI sequence evolution could be much faster than other genes and still be consistent with vertical evolution, since rates of evolution can vary extensively in different lineages. The inconsistency, manifesting as evidence for likely horizontal transfer, would be rates that are much slower than an established minimum amount (determined either by geological or relative timings).

Tree topology places the first gap as occurring between closest cyanobacterial PADI homologues and PADIs in the earliest diverging metazoa (Figure 3.3). Analysing these orthologues in isolation, a phylogenetic tree was made using all cyanobacterial PADI sequences with a representative set of metazoan PADIs (Figure 3.7). Known cyanobacterial clades derived from the latest multi-gene predictions of cyanobacterial species trees were used to annotate the PADI paralogue tree. Notably, the PADI gene tree mirrors the known pattern of vertical evolution of

cyanobacterial clades. The closest ancestral PADI homologues to the metazoan sequences are found in two clades of late-diverging cyanobacterial species: the *NX* clade (containing Nostocales sensu lato + others) and the *SPM* clade (containing Synechocystis + Pleurocapsa + Microcystis)[54]. Subsequently, rates of PADI evolution were analysed using PADI genomic sequences from these *NX* and *SPM* cyanobacterial clades and from metazoa. These were obtained from the PATRIC database.

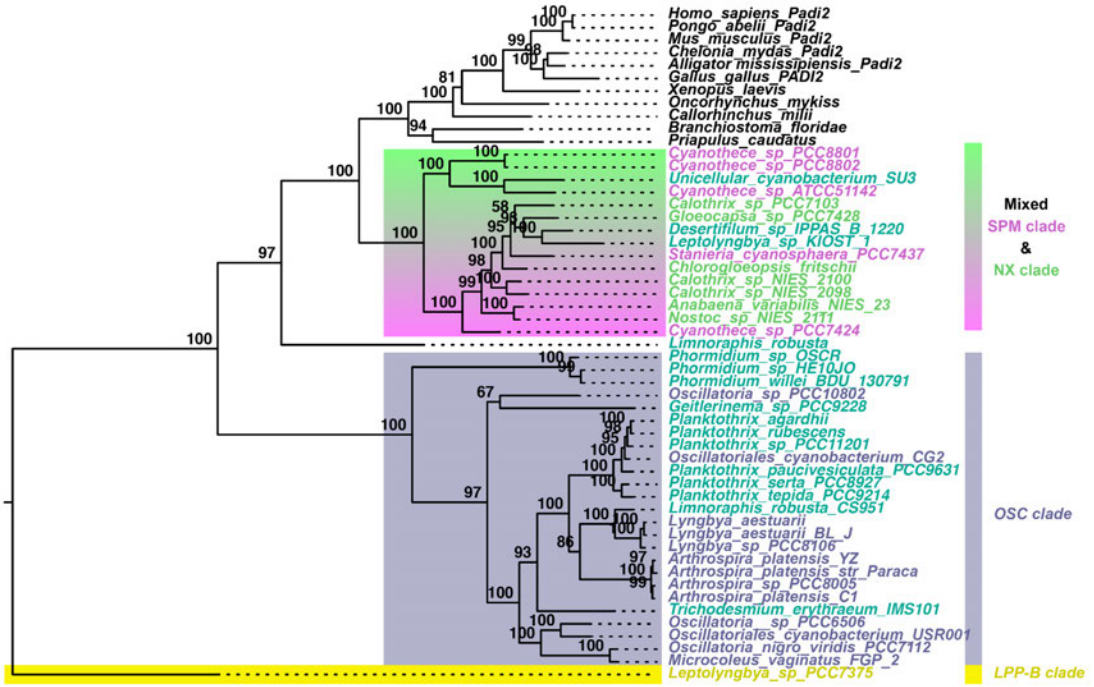


Figure 3.7: Phylogenetic analysis of metazoan and cyanobacterial PADIs. All putative PADIs from cyanobacteria were aligned with a subsampled set of metazoan sequences using MAFFT L-ins-I and singly aligning columns were removed. IQTree was used to produce a maximum likelihood phylogenetic tree using the WAG empirical rate matrix with 5 categories of rate variation under the FreeRate model (LG +R8) with base frequencies counted from the alignment as determined by ModelFinder. Ultrafast Bootstrap 2 with 1000 replicates, SH-like aLRT with 1000 replicates and aBayes parametric tests were used to assess node support. The tree is shown unrooted with Ultrafast Bootstrap 2 node support labelled on the tree. Colours and names of clades are used as in *Uyeda et al*. 2016[54]. Species that were not analysed in *Uyeda et al*. 2016 are coloured in cyan and metazoan sequences in black.

Firstly, analysing the clade of cyanobacterial and animal PADI sequences using a Bayesian phylogenetic approach allowed their divergence time to be predicted using known fossil ages of metazoans[55-57] as calibrations under a strict molecular clock model (Section 2.5.4.1). This yielded an estimate of approximately 1 billion years (Figure 3.8) for the age of their last common ancestor, far younger than the 3.35-4.52 billion years known to separate bacteria and eukarya (Figure 3.8)[53]. Divergence times of vertebrate and cyanobacterial PADIs were then estimated using several relaxed clock models (UCLN, UCED, random local clocks)[57] (Figure 3.8B). These relaxed clocks increased the uncertainty in the estimate but, in all three cases, estimated an even more recent mean divergence time. Under all approaches, the divergence times were not congruent with the geologically-defined divergence ($p < 10^{-8}$) (Figure 3.8B). These divergence time estimates are instead consistent with a horizontal acquisition more recently than that of the mitochondria and are approximately dated to the divergence time of the last common ancestor of the earliest diverging metazoa that possess a PADI.

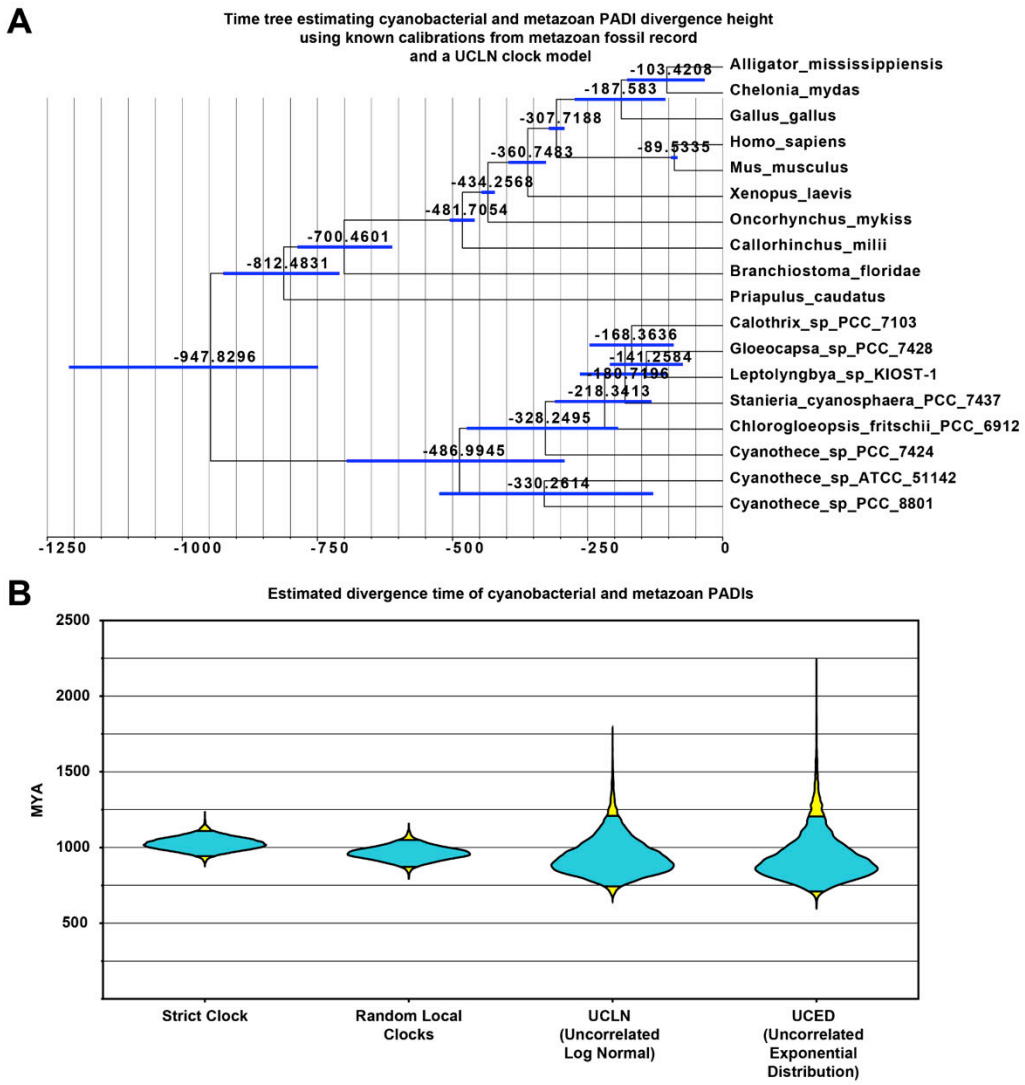


Figure 3.8: Estimated divergence time of cyanobacteria and metazoa based on PADI sequences with respect to geologically defined constraints from the fossil record. **A:** Metazoan sequences and *SPM/NX* clade of cyanobacterial sequences taken from the PATRIC database are used for Bayesian phylogenetic analysis. Bayesian phylogenetic analysis was performed using BEAST under the uncorrelated lognormal (UCLN) clock model using a calibrated Yule model as the tree prior (Chapter 2.5.4). Divergence times from the fossil record are used as normally distributed priors for six different nodes from metazoa. The marginal posterior distribution of the age of the root of the whole tree is used to estimate the divergence time; all nodes are labelled with the 95% credible interval for the marginal posterior distributions of the node ages. **B:** Estimated divergence times for running the analysis under different clock models (strict clock, random local clocks, UCLN and UCED), with the marginal posterior distribution of the age of the root of the whole tree given as a violin plot, where the 95% credible interval is given in cyan and the limits coloured in yellow.

Secondly, analysing a large number of the most conserved proteins in life (ribosomal proteins, essential metabolic enzymes, chaperones), taken from species that bridge the closest PADI homologues (in cyanobacteria and in metazoa), approximates a mean minimum extent (the difference in units of bitscore density) of accumulated genetic divergence occurring in both lineages over this time scale. To do this analysis, the bitscore density of the similarity of the cyanobacterial homologue to the human sequence $\Delta bitscore_{Cy\text{-}Hu}$ (AC+AH) and the bitscore density of the similarity of the branchiostomal homologue to the human sequence $\Delta bitscore_{Br\text{-}Hu}$ (XB+XH) were calculated (Figure 3.9A). A measure of the total accumulated genetic divergence between late-diverging cyanobacteria (*Cyanothece spp.*) and the last common ancestor of *Branchiostoma spp.* and *Homo sapiens* was calculated by subtracting $\Delta bitscore_{Br\text{-}Hu}$ from the $\Delta bitscore_{Cy\text{-}Hu}$, as illustrated in Figure 3.9A. This was performed on 26 highly conserved proteins and the mean was calculated (Figure 3.9B). The distribution of calculated accumulated genetic divergence for these different highly conserved proteins did not deviate significantly from a normal distribution (Shapiro-Wilk test). As a positive control for the hypothesized horizontal trajectory in Figure 3.9A, I then analysed proteins of likely endosymbiont gene transfer (EGT) origin (19 sequences) as well as proteins encoded in the mitochondrial genomes (10 sequences). A full list of analysed proteins is provided in Section 2.5.4.2. Since mitochondrial and EGT-derived proteins were acquired more recently than the LUCA, the mean of the total accumulated sequence change for each protein is expected to be much lower than that for vertically transferred genes (Figure 3.9A). It was therefore hypothesized that they may mimic more closely the extent of accumulated genetic divergence that would be expected for an anciently horizontally transferred gene (such as is hypothesized for the PADI gene). As expected, EGT and mitochondrially encoded proteins have an average accumulated genetic divergence which is significantly lower than that of vertically acquired proteins (Figure 3.9B). Results plotted in Figure 3.9B and 3.9C revealed that the total accumulated genetic divergence for PADI sequences, and assessed over exactly the same timescale, falls 6

standard deviations below that calculated for vertically transferred protein sequences. PADIs show less sequence change than all proteins individually analysed over this timescale and less even than ribosomal RNA (Section 2.5.4.3). Indeed, they fall 2 standard deviations even below the mean of EGT candidate genes or the mean of genes derived from the mitochondrial genome (Figure 3.9B and C). In a model of vertical descent, PADIs would therefore be under greater constraint than any sequence known in life[58].

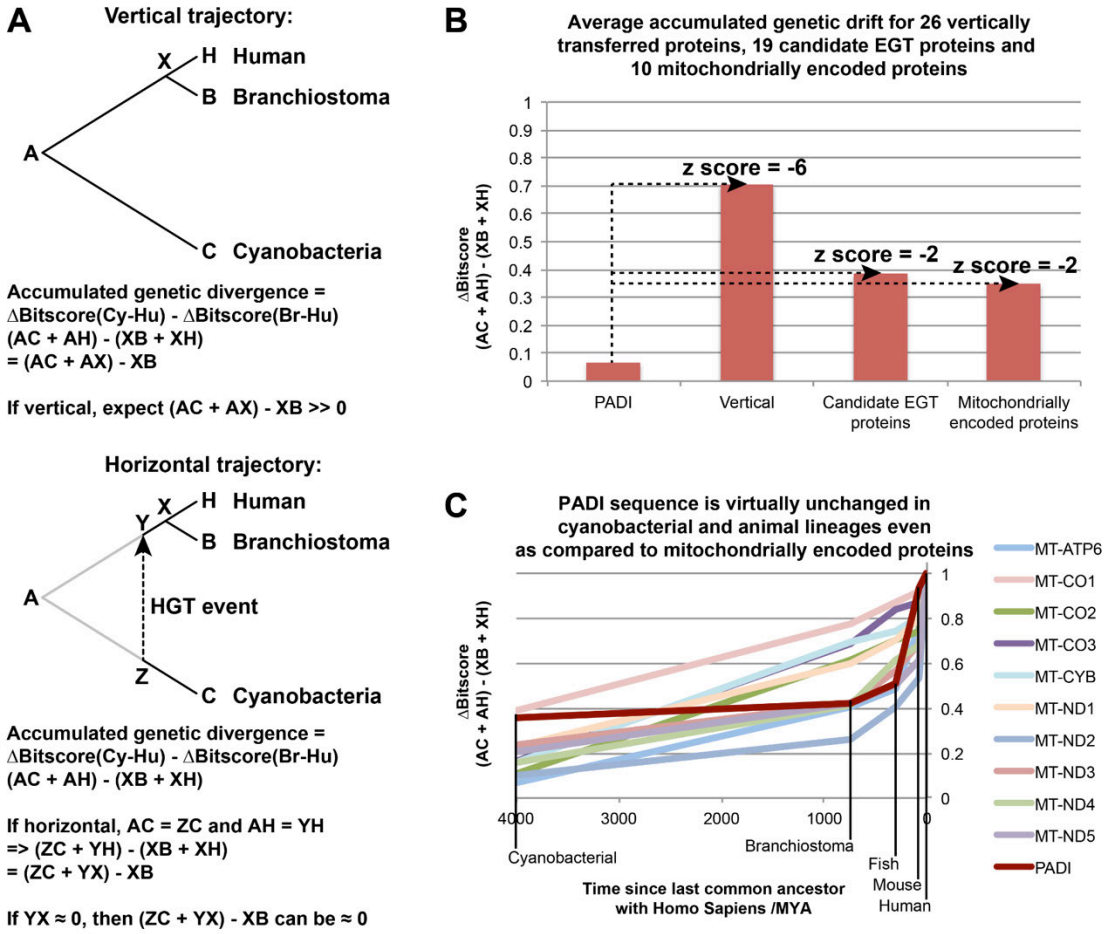


Figure 3.9: Analysis of the Accumulated Genetic Divergence of PADIs compared to very well conserved proteins. **A:** Schematic of how the Accumulated Genetic Divergence (AGD) of a given protein between its homologues in *Homo sapiens*, *Branchiostoma spp*. and *Cyanothece spp*. is calculated. The bitscore density of the similarity of the cyanobacterial homologue to the human sequence $\Delta bitscore_{Cy\text{-}Hu}$ (AC+AH) and the bitscore density of the similarity of the branchiostomal homologue to the human sequence $\Delta bitscore_{Br\text{-}Hu}$ (XB+XH) are calculated. In a vertical scenario the AGD, given by (AC+AX) – XB, will be much greater than zero. In a horizontal scenario, (ZC+YX) may be approximately equal to XB and so the

AGD, given by (ZC + YX) – XB may be close to zero. **B:** The AGD was calculated for 26 vertically transferred proteins, 19 candidate EGT proteins, and 10 proteins encoded in the mitochondrial genome and the mean was calculated and plotted against the AGD for PADI proteins. **C:** The difference in bitscore density between mitochondrial protein sequences from various species and their human counterpart was calculated and plotted against the age of the last common ancestor for each pair (human and mouse, human and fish, human and branchiostoma, human and cyanobacteria). This was similarly calculated for PADIs from the same species pairs and plotted on the same graph.

To contextualize these data, in absolute terms, the similarity of cyanobacterial and branchiostomal PADIs to human PADIs is almost identical: 70.20% vs 70.90% respectively by pairwise amino acid similarity to the human sequence. In a model of vertical evolution, however, a much greater amount of time has elapsed since the cyanobacterial and human genes have shared a last common ancestor than the genes from the other species pair (branchiostoma and humans). Under the assumption of horizontal transfer, the explanation for the observation of such little change in sequence is more mundane. An HGT event from late-diverging cyanobacteria to a last common ancestor within the animal lineage, although ancient, would have occurred much more recently than the LUCA and somewhat more recently than the mitochondrion too (Figure 3.9A, Figure 3.12). HGT can therefore fully account for the slow rates of evolution observed even compared to the rate of evolution of EGT candidates and mitochondrially encoded proteins. This evidence (Figure 3.9) therefore is complementary to and independent from the estimates derived from divergence time predictions based on the fossil record (Figure 3.8).

Although rates of evolution may differ (even extensively) between different lineages (heterotachy), a minimal amount of nearly neutral genetic divergence nonetheless accumulates over evolutionary timescales in all lineages, even in the best-conserved genomic sequences in life[58]. Under the assumption of vertical descent, the observed PADI sequence changes are anachronistically low relative to other vertically acquired genes (6 standard deviations lower). In combination, the rate of PADI sequence evolution

across the phylogeny, calibrated using timings from the metazoan fossil record under various relaxed molecular clock models[57], requires a common ancestry of cyanobacterial PADIs and animal PADIs that is considerably more recent (at approximately 1 billion years ago; Figure 3.8) than the well-defined last common ancestry of cyanobacteria and eukarya (approximately 4 billion years ago)[53].

3.2.6 Catalytic activity and calcium dependence of cyanobacterial PADI

Given the high conservation of cyanobacterial PADI to metazoan PADIs, including all necessary catalytic residues and calcium binding residues, I hypothesized that it was likely that it might be catalytically active. To assess whether the cyanobacterial protein is capable of citrullination, a recombinant version of the three-domain PADI from *Cyanothece sp. 8801* (referred to as "cyanoPADI") was prepared. DNA encoding the three-domain cyanoPADI sequence was synthesized commercially by Thermo Fisher GeneArt. This sequence, and the sequence for human PADI4, were sub-cloned into a bacterial expression vector with an N-terminal Glutathione S-Transferase-6xHistidine (GST-His) tag (*sub-cloning was performed by Gavriil Gavriilidis and Abigail Wilson*). The GST tag helps to solubilize the mammalian protein and purification with the His tag provides a good yield from affinity purification. N-terminal tags (Flag, 6xHis, and the larger GST) have been shown previously not to affect the activity of PADI4 *in vitro*[59]. The two proteins were purified under identical conditions from bacterial cell lysates. GST-His-CyanoPADI and GST-His-humanPADI4 were then used directly in *in vitro* citrullination assays (Chapter 2.2.6.1). Detection of proteins was carried out by Western blot, and loading assessed using an antibody to the GST tag and to two housekeeping proteins (NPM1, GAPDH). Mouse cell lysates were used to provide a broad range of possible protein substrates for possible citrullination.

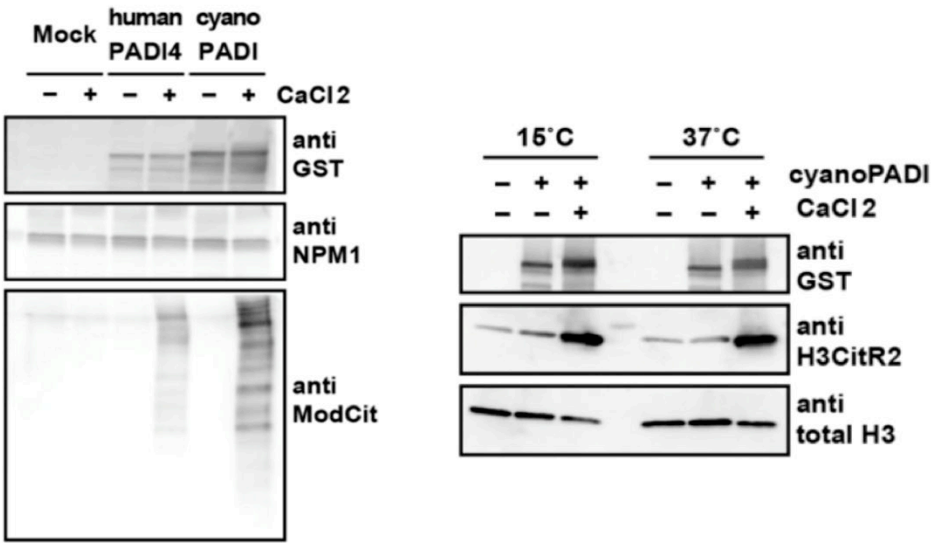


Figure 3.10: Cyanobacterial PADI enzyme from *Cyanothece sp. 8801* (cyanoPADI) is catalytically active *in vitro*. Immunoblot analysis of citrullination assays using recombinant GST-His tagged enzymes on **A:** mouse embryonic stem cell lysates and **B:** recombinant human histone H3 with detection by Mod-Cit and H3CitR2 respectively. NPM1 and total histone H3 are presented as loading controls. Recombinant cyanobacterial protein production was performed with Gavriil Gavriilidis. Data shown in A are representative of n = 3; the Mod Cit detection was performed once. Data shown in B are representative of n = 2.

Analogously to the human enzyme, cyanoPADI citrullinated multiple proteins in mouse cell lysates (Figure 3.10A), as detected by Mod-Cit, which is an antibody to chemically modified citrulline residues (Chapter 2.1.16). CyanoPADI, like human PADI4, showed an absolute dependence on calcium for activity (Figure 3.10). This demonstrates that the calcium-dependent regulation found in mammalian PADIs is also a feature of the ancestral cyanobacterial protein and provides evidence that the conserved calcium-binding sites, used in the evolutionary analysis as signifiers of synapomorphy, are functional. Despite the absence of histones in bacteria, cyanoPADI catalysed citrullination of arginine 2 on the N-terminal tail of recombinant human histone H3 (Figure 3.10B), which is a known target of mammalian PADI4. Thus cyanoPADI is a bona fide calcium-dependent peptidylarginine deiminase with sufficient similarity or promiscuity to catalyse citrullination of mammalian target substrates. It was noticeable that a slight band shift to apparent higher molecular weight occurred to the tagged GST-

His proteins in the activated (calcium-treated) conditions. This may be due to a modification to the protein, which is most likely due to autocitrullination as has been shown for PADI4[60], but could also be citrullination of the affinity tag, or less likely an activated conformation of the protein behaving differently under denatured protein electrophoresis conditions. The assay was repeated at 15˚C and 37˚C and activity was reproduced, showing the cyanobacterial enzyme is active at a physiologically relevant temperature in the ocean (Figure 3.10B). A variety of other metal ions did not substitute for activation by calcium ions under the same conditions in a pilot experiment (data not shown).

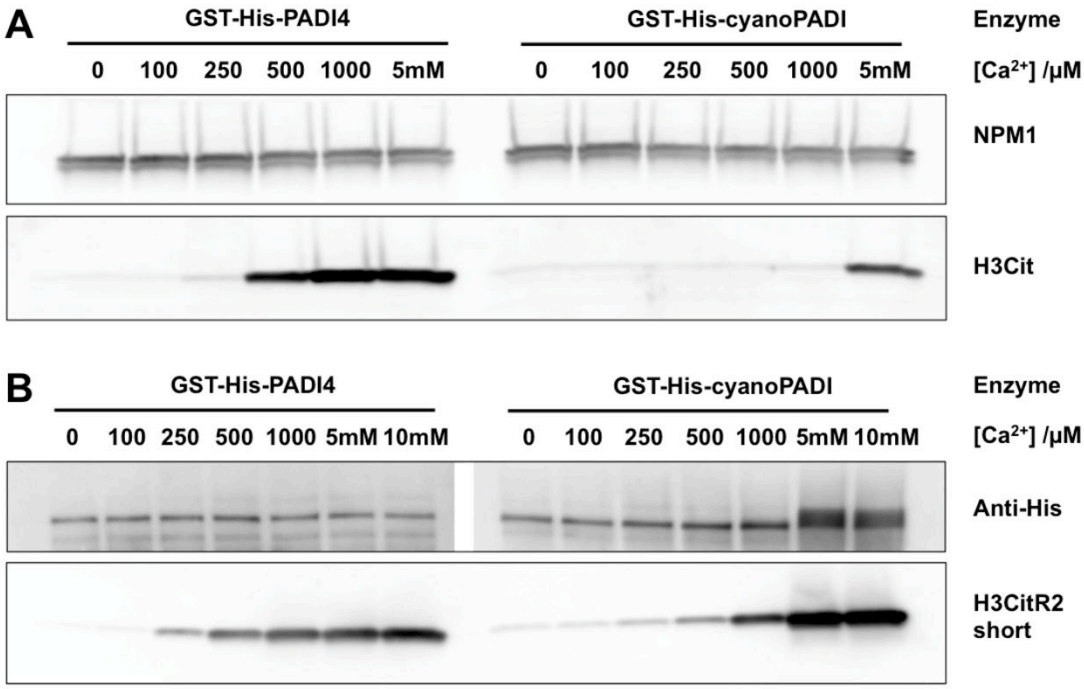


Figure 3.11: Cyanobacterial PADI enzyme from *Cyanothece sp. 8801* shows a different Ca2+ dependency *in vitro*. Immunoblot analysis of citrullination assays using recombinant GST-His tagged enzymes on **A:** mouse embryonic stem cell lysates for which data shown are representative of n = 3, and **B:** recombinant human histone H3 with detection by H3CitR2 for which data are shown from a single preliminary experiment. NPM1 and anti-His are presented as loading controls.

Finally, given the high calcium requirement of the mammalian enzyme, the citrullination assay was repeated at a wide range of calcium concentrations

to test whether there was any variation in calcium dependence between the human and cyanobacterial PADI enzymes (Figure 3.11). Human PADI4 showed some activity in $CaCl_2$ concentrations as low as ~250 μM *in vitro* on Arg2 of histone 3 (Figure 3.11A) with an incubation time of 30 minutes. However, cyanoPADI showed no activity on the same substrate even at 1 mM $CaCl_2$, with activity demonstrated only at $CaCl_2$ concentrations greater than 5 mM (Figure 3.11A). It was then tested in a pilot experiment whether the stricter Ca^{2+} dependence could also be observed on a direct recombinant substrate in the absence of mouse cell lysate proteins (Figure 3.11B). This is important because it is possible that there is a cofactor present in mouse cell lysates that can activate the human enzyme, but not the cyanobacterial PADI. This pilot experiment showed the same pattern (Figure 3.11B). Robust activation of the cyanobacterial PADI occurred only at 1000 μM $CaCl_2$, compared to PADI4 activation occurring as low as 250 μM $CaCl_2$. This will need to be repeated at exactly the same time point of incubation as used for the lysate assay as there is discrepancy in the minimum concentration of Ca^{2+} ions required in longer incubations. Testing this comparison at different incubation times will assist in making firm comparisons.

Nonetheless this work shows that cyanoPADI has a stricter requirement for calcium than the mammalian PADI proteins and in preliminary terms these experiments indicate this effect is observed independently from the presence of cofactors in the cellular lysate. This suggests the intrinsic affinity for calcium may have evolved to be higher in the vertebrate lineage. The intrinsic affinity of human PADI4 for calcium ions may therefore be higher than for cyanoPADI. Testing for the full calcium dependence on cyanobacterial PADI activity on a recombinant substrate in further more quantitative biochemical experiments will therefore be very interesting (such as obtaining Hill constants and K_m values for Ca^{2+})[61].

3.3 Discussion of horizontal gene transfer

HGT is established for its importance in prokaryotes[14], and although suggested examples in eukaryotes are increasingly abundant in the literature, the strength of evidence varies enormously between individual cases[62,63]. HGT examples are therefore regularly contested. Two extremes of evidence for HGT into eukaryotes exist. On the one hand, recent examples of HGT into a very limited number of species can be relatively convincing if contamination can be definitively ruled out[64]. HGT is argued by parsimony because the alternative of vertical descent requires near complete gene losses of the missing ancestral homologues across the tree of life. On the other hand, examples of ancient HGT of a gene showing wider taxonomic distribution after transfer are unlikely to be caused by contamination, but may more easily be products of vertical transfer that are confounded by gene loss and rate variation in different lineages. In such examples, if phylogenetic analysis of a single gene alignment shows multiple instances of gene-species discordance, commonly used arguments for a vertical alternative to HGT become very convoluted[23,65,66]. The caveat remains, however, that this type of analysis is reliant on the assumed accuracy of phylogenetic inference of a single gene over long spans of evolutionary time. Neither case is conclusive, and unsurprisingly, many much less 'conclusive' cases lie somewhere between these extremes. In addition, many individual HGT candidates into eukaryotes may be explained by the related phenomenon of endosymbiont gene transfer (EGT), such as genes of plastid or mitochondrial origin, which were subsequently acquired in the nucleus[66-68]. At some point, evidence for HGT almost inevitably descends into a discussion of parsimony. This unfortunately has had the effect of reducing the relative strength of cases of eukaryotic HGT down to one's prior perceptions of the relative likelihood of HGT weighed against alternatives such as independent gene losses[20,69]. This has brought into question once again whether HGT has ever occurred into eukaryote genomes[70-73]. As Salzberg writes[20]:

"My re-examination here suggests that HGT is very rare rather than widespread in vertebrate genomes, and that every hypothesized HGT event

needs to be subjected to careful scrutiny. […] Because HGT is such an unlikely event, the results of automated searches should be subjected to individual, close scrutiny with an eye toward explaining them through more mundane processes before concluding that these anomalies represent novel biological discoveries."

This chapter therefore attempts to address these concerns through consilience of evidence, presenting multiple independent methods that converge on horizontal transfer in the instance of the PADI gene.

Here, three types of analyses of PADI sequence evolution were undertaken. The first two analyses – of phylogenetic tree topology and synapomorphies – are consistent with vertical descent (but with very large numbers of independent gene losses) or with horizontal transfer from an ancestral cyanobacterium into an early metazoan. Synapomorphic evidence that confirms the phylogenetic topology at the level of the protein sequence or a protein domain[74-76] is particularly important: it is robust both to convergent evolution[52] and phylogenetic artifacts derived from the analysis of a single gene over long periods of evolutionary time. Results from the third, molecular clock, analysis are not easily explained by vertical descent. This is because this scenario requires the rate of PADI sequence evolution to have slowed sharply in evolution twice, in lineages leading to metazoa and in cyanobacterial lineages, while being absent from all other lineages. This is measured to be anachronistic both in absolute terms with respect to fossil calibrations using molecular clocks, as well as in relative terms as compared to other vertically transferred proteins and even to mitochondrially encoded proteins.

Ruling out EGT can often be very complex; it is conceivable that after a gene from the mitochondrion was transferred to the nucleus it could have duplicated and then undergone differential loss and rate variation in different lineages that lead to modern genomes. In the case of *PADI* genes, EGT can be ruled out as two distinct eukaryotic *PADIs* can be identified and shown to be homologous to different bacterial *PADIs*. This is evidenced by the

phylogenetic topology, and confirmed by synapomorphies in the sequence that rule out convergent evolution or phylogenetic artifacts. This is consistent with analyses of the rate of *PADI* evolution, which point to a more recent acquisition than the mitochondrion.

The direction of HGT can be seen to be from cyanobacteria to metazoa and not in reverse. This is shown by phylogenetic analysis and confirmed by the synapomorphic evidence. In addition, two additional pieces of evidence are found: firstly, strong support is found for bacterial outgroup PADIs to both types of eukaryotic sequence (Figure 3.2) and secondly, the cyanobacterial *PADI* gene tree mirrors the expected species tree (Figures 3.8 and 3.10)[54]. A three-domain PADI therefore emerges only after *SPM* and *NX* clades diverge (Figure 3.12).

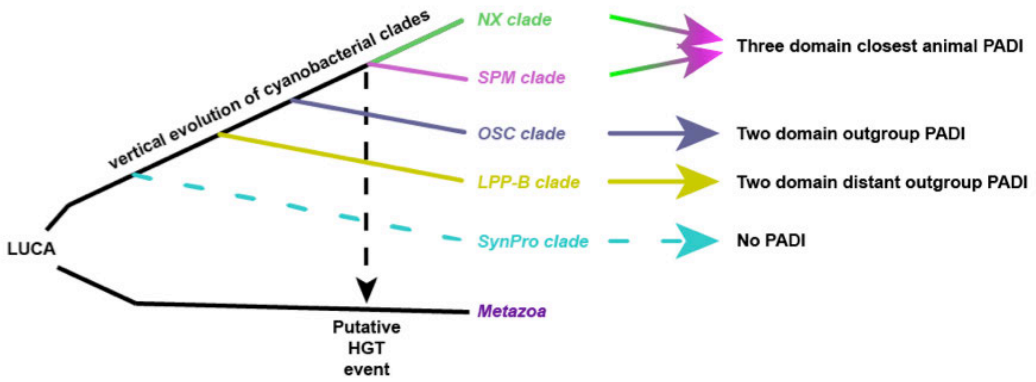


Figure 3.12: Schematic showing the proposed HGT event from cyanobacteria to metazoa as occurring in a last common ancestor of the *NX/SPM* clades of cyanobacteria, which possess a three-domain PADI and transferring to an ancient last common ancestor in the metazoan lineage.

3.4 Discussion

PADIs, therefore, provide an unusually compelling case of ancient horizontal transfer of a bacterial gene into eukaryotes and the acquisition of a new catalytic function. Although found in fungi and in animals, each PADI type was acquired independently from different bacterial sources, firstly by

animals from cyanobacteria and secondly by fungi from actinobacteria. In the case of the animal *PADI* gene, it was lost in many lineages shortly after transfer and only maintained widely in the vertebrate lineage, persisting into the human genome. Fish have a single gene, but duplications down the vertebrate lineage have resulted in five tandem repeated paralogues in mammals, each with controlled tissue-specific expression. The timing of transfer (neoproterozoic: 1000-542MYA) is coincident firstly with the presence of marine nitrogen fixing cyanobacteria with specialized arginine catabolic pathways[77] that may have lead to the evolution of the cyanobacterial PADI, and secondly with the emergence of metazoa in the oceans[78-80].

The newly acquired catalytic function (i.e. citrullination) was co-opted in animals for very diverse functions in specific cellular contexts, ranging from neutrophils to pluripotent stem cells, and from oligodendrocytes to keratinocytes[4,59,81]. This functional diversity and locational specificity raises an interesting question as to the selective pressure for maintenance of PADIs after transfer. It is notable therefore that PADI6-knockout mice and inactivating mutations in human PADI6 cause female infertility[82-85]. To address why this product of horizontal transfer was maintained in the chordate lineage but not in multiple others, it would be interesting to know whether PADI2-knockout zebrafish are viable since fish possess only a single PADI paralogue. The roles for PADIs in classically animal-specific processes such as the maturation of myelin and hair follicles are particularly interesting from this evolutionary standpoint.

As to the function in bacteria, literature searches of ribosomally synthesized and post-translationally modified peptide (RiPP) natural products revealed a recent paper that identified citrullinated peptides in bacteria, including in citrullassin A from *Streptomyces albulus*[86]. In this paper, a candidate citrullinating enzyme could not be identified. It is very interesting therefore that the same *Streptomyces albulus* does possess a putative peptidyl

arginine deiminase (which I identified from HMMER searches to the PAD_C domain, Figure 3.1A and was analysed in Figure 3.2). This therefore provides a concrete example of PADI activity in bacteria and a proper function for the putative bacterial PADI. A similar search of cyanobactins (small heavy post-translationally modified cyclic peptides produced by cyanobacteria) did not identify citrullinated residues in already published papers but this appears to be a likely possibility. In addition, a putative cyanophycinase was identified as being contained within the same operon as the putative PADI in *Cyanothece sp. 51142*. Cyanophycin is a polyarginine-polyasparagine polymer used for cyanobacterial energy storage. It would be very interesting if this were modified to citrulline in cyanobacteria. PADIs are known to be highly efficient enzymes in mammalian cells and multiply modify myelin basic protein (up to 80% in a Marburg variant of MS[87]) as well as keratins. It is plausible that the multiple arginine residues in cyanophycin may provide multiple sites for citrullination in cyanobacteria and a similar mode of activity.

Possibly related are considerations about the stringent PADI regulation observed in mammals. Calcium-binding residues are common to all PADI homologues and a similar mechanism of ordered calcium regulation is also most likely fully conserved in the closest ancestral cyanobacterial homologues. Fungal and actinobacterial sequences are likely to be regulated somewhat differently, but also in a calcium dependent manner as they still conserve at least three different sites. It would clearly be interesting to test this specifically by synthesizing a fungal and actinobacterial PADI homologue and generating point mutants. It is notable that human PADI4 is active at tenfold lower calcium concentrations than the ancestral cyanobacterial protein using mouse cell lysates. It is possible that a cofactor present in mouse cell lysates could support the lower calcium requirement of human PADI4 than cyanobacterial PADI in these assays. Alternatively, as $CaCl_2$ levels in the ocean are at 10 mM, this would support extracellular activity of the cyanobacterial protein if it were secreted. There is evidence for the

secretion or extracellular activity of PADI2 and PADI4 by neutrophils as well as for pPAD by *Porphyrmonas gingivalis*[88-90]. Whether an animal-specific evolutionary context may have resulted in the physiological regulation in human cells is an important additional possibility. The intracellular roles in mammals are likely to have derived from evolving a higher affinity to calcium. In mammals, there is likely still to be a requirement for an additional activating cofactor or a mechanism to support very high local calcium elevation in viable cells such as pluripotent stem cells, as even intracellular calcium concentrations spikes result in peaks of no more than ~1 μM concentration, with resting levels measured to be below 100 nM. If the cyanobacterial protein is active intracellularly, then the even higher calcium stringency of the cyanobacterial protein offers an analogous mystery.

A final reflection, in light of the discovery of bona fide bacterial PADIs, relates to the lack of discovery of a 'decitrullinase' enzyme to date. 'Decitrullination' activity on free citrulline has precedent in nature from two enzymes in the urea cycle; citrulline is converted to argininosuccinate first by argininosuccinate synthase (ASS), and the succinyl group subsequently cleaved to free arginine by argininosuccinate lyase (ASL). The horizontal origin of animal PADIs complicates the possibility of finding a reverse enzyme (or enzymes) that may only have arisen or been maintained in species which possess citrullinating catalytic competency. At the very least, a search for a decitrullinase, if it does exist, should be expanded to include bacterial and fungal proteins. Alternatively models of protein degradation to explain the dynamic nature of arginine methylation could be invoked for a similar consideration of the removal and therefore dynamic nature of citrullination[91], if there is no animal decitrullinase enzyme.

3.5 References for Chapter 3

1. Balandraud, N. *et al.* A rigorous method for multigenic families' functional annotation: the peptidyl arginine deiminase (PADs) proteins family example. *BMC Genomics* **6,** 153 (2005).
2. György, B., Tóth, E., Tarcsa, E., Falus, A. & Buzás, E. I. Citrullination: A posttranslational modification in health and disease. *The International Journal of Biochemistry & Cell Biology* **38,** 1662–1677 (2006).
3. Wang, S. & Wang, Y. Peptidylarginine deiminases in citrullination, gene regulation, health and pathogenesis. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1829,** 1126–1135 (2013).
4. Nicholas, A. P. & Bhattacharya, S. K. *Protein deimination in human health and disease*. (2014).
5. Shirai, H., Blundell, T. L. & Mizuguchi, K. A novel superfamily of enzymes that catalyze the modification of guanidino groups. *Trends in Biochemical Sciences* **26,** 465–468 (2001).
6. Linsky, T. & Fast, W. Mechanistic similarity and diversity among the guanidine-modifying members of the pentein superfamily. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **1804,** 1943–1953 (2010).
7. Arita, K. *et al.* Structural basis for Ca2+-induced activation of human PAD4. *Nature Structural & Molecular Biology* **11,** 777–783 (2004).
8. Slade, D. J. *et al.* Protein arginine deiminase 2 binds calcium in an ordered fashion: implications for inhibitor design. *ACS Chem. Biol.* **10,** 1043–1053 (2015).
9. McGraw, W. T., Potempa, J., Farley, D. & Travis, J. Purification, Characterization, and Sequence Analysis of a Potential Virulence Factor from Porphyromonas gingivalis, Peptidylarginine Deiminase. *Infect. Immun.* **67,** 3248–3256 (1999).
10. Carolina Touz, M. *et al.* Arginine deiminase has multiple regulatory roles in the biology of Giardia lamblia. *J Cell Sci* **121,** 2930–2938 (2008).
11. Söding, J. Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21,** 951–960 (2005).
12. Holm, L. & Rosenström, P. Dali server: conservation mapping in 3D. *Nucl. Acids Res.* **38,** W545–9 (2010).
13. Montgomery, A. B. *et al.* Crystal structure of Porphyromonas gingivalis peptidylarginine deiminase: implications for autoimmunity in rheumatoid arthritis. *Ann Rheum Dis* **75,** 1255–1261 (2016).
14. Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405,** 299–304 (2000).
15. Goldenfeld, N. & Woese, C. Biology's next revolution. *Nature* **445,** 369–369 (2007).
16. Consortium, I. H. G. S. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
17. Salzberg, S. L., White, O., Peterson, J. & Eisen, J. A. Microbial genes in the human genome: Lateral transfer or gene loss? *Science* **292,** 1903–1906 (2001).
18. Crisp, A., Boschetti, C., Perry, M., Tunnacliffe, A. & Micklem, G. Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes. *Genome Biol.* **16,** 50 (2015).
19. Jensen, L., Grant, J. R., Laughinghouse, H. D. & Katz, L. A. Assessing the effects of a sequestered germline on interdomain lateral gene transfer in Metazoa. *Evolution* **70,** 1322–1333 (2016).
20. Salzberg, S. L. Horizontal gene transfer is not a hallmark of the human genome. *Genome Biol.* **18,** 85 (2017).
21. Ku, C. & Martin, W. F. A natural barrier to lateral gene transfer from prokaryotes to eukaryotes revealed from genomes: the 70 % rule. *BMC Biol.* **14,** (2016).
22. Stanhope, M. J. *et al.* Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. *Nature* **411,** 940–944 (2001).
23. Andersson, J. O., Doolittle, W. F. & Nesbø, C. L. Are There Bugs in Our Genome? *Science* **292,** 1848–1850 (2001).
24. Roelofs, J. & Van Haastert, P. Genomics - Genes lost during evolution. *Nature* **411,** 1013–1014 (2001).

25. DeFilippis, V. & Villarreal, L. P. Lateral gene transfer or viral colonization? *Science* **293,** 1048–1048 (2001).
26. Salzberg, S. L. & Eisen, J. A. Lateral gene transfer or viral colonization? Response. *Science* **293,** 1048–1048 (2001).
27. Willerslev, E. *et al.* Contamination in the draft of the human genome masquerades as lateral gene transfer. *DNA Seq.* **13,** 75–76 (2002).
28. Genereux, D. P. & Logsdon, J. M. Much ado about bacteria-to-vertebrate lateral gene transfer. *Trends Genet.* **19,** 191–195 (2003).
29. Boschetti, C. *et al.* Biochemical Diversification through Foreign Gene Expression in Bdelloid Rotifers. *PLoS Genet.* **8,** e1003035 (2012).
30. El-Sayed, A. S. A. *et al.* Biochemical characterization of peptidylarginine deiminase-like orthologs from thermotolerant Emericella dentata and Aspergillus nidulans. *Enzyme Microb. Technol.* **124,** 41–53 (2019).
31. Huerta-Cepas, J. *et al.* eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucl. Acids Res.* **44,** D286–D293 (2016).
32. Jones, D. T., Taylor, W. R. & Thornton, J. M. The rapid generation of mutation data matrices from protein sequences. *Bioinformatics* **8,** 275–282 (1992).
33. Whelan, S. & Goldman, N. A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. *Mol. Biol. Evol.* **18,** 691–699 (2001).
34. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Haeseler, von, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* **14,** 587–589 (2017).
35. Lopez, P., Casane, D. & Philippe, H. Heterotachy, an important process of protein evolution. *Mol. Biol. Evol.* **19,** 1–7 (2002).
36. Felsenstein, J. Cases in which Parsimony or Compatibility Methods will be Positively Misleading. *Syst Biol* **27,** 401–410 (1978).
37. Wang, H.-C., Susko, E., Spencer, M. & Roger, A. J. Topological estimation biases with covarion evolution. *J Mol Evol* **66,** 50–60 (2008).
38. Ronquist, F. *et al.* MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Syst Biol* **61,** 539–542 (2012).
39. Quang, L. S., Gascuel, O. & Lartillot, N. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* **24,** 2317–2323 (2008).
40. Lartillot, N. & Philippe, H. A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process. *Mol. Biol. Evol.* **21,** 1095–1109 (2004).
41. Huelsenbeck, J. P. Is the Felsenstein zone a fly trap? *Syst Biol* **46,** 69–74 (1997).
42. Philippe, H. & Laurent, J. How good are deep phylogenetic trees? *Current Opinion in Genetics & Development* **8,** 616–623 (1998).
43. Lartillot, N., Brinkmann, H. & Philippe, H. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol. Biol.* **7,** (2007).
44. Hoang, D. T., Chernomor, O., Haeseler, von, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **35,** 518–522 (2018).
45. Felsenstein, J. Confidence Limits on Phylogenies: an Approach Using the Bootstrap. *Evolution* **39,** 783–791 (1985).
46. Strimmer, K. & Rambaut, A. Inferring confidence sets of possibly misspecified gene trees. *Proc. Biol. Sci.* **269,** 137–142 (2002).
47. Shimodaira, H. An Approximately Unbiased Test of Phylogenetic Tree Selection. *Syst Biol* **51,** 492–508 (2002).
48. Susko, E. Tests for two trees using likelihood methods. *Mol. Biol. Evol.* **31,** 1029–1039 (2014).
49. Finn, R. D. *et al.* HMMER web server: 2015 update. *Nucl. Acids Res.* **43,** W30–W38 (2015).
50. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology* **292,** 195–202 (1999).
51. Sonnhammer, E. L. & Hollich, V. Scoredist : A simple and robust protein sequence distance estimator. *BMC Bioinformatics* **6,** 1–8 (2005).

52. Doolittle, R. F. Convergent evolution: the need to be explicit. *Trends in Biochemical Sciences* **19,** 15–18 (1994).
53. Betts, H. C. *et al.* Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. *Nature Ecology & Evolution 2018 2:5* **2,** 1556–1562 (2018).
54. Uyeda, J. C., Harmon, L. J. & Blank, C. E. A Comprehensive Study of Cyanobacterial Morphological and Ecological Evolutionary Dynamics through Deep Geologic Time. *PLoS ONE* **11,** (2016).
55. Bouckaert, R. *et al.* BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLOS Computational Biology* **10,** e1003537 (2014).
56. Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. Relaxed Phylogenetics and Dating with Confidence. *Plos Biol* **4,** e88 (2006).
57. Drummond, A. J. & Suchard, M. A. Bayesian random local clocks, or one rate to rule them all. *BMC Biol.* **8,** (2010).
58. Isenbarger, T. A. *et al.* The Most Conserved Genome Segments for Life Detection on Earth and Other Planets. *Orig Life Evol Biosph* **38,** 517–533 (2008).
59. Christophorou, M. A. *et al.* Citrullination regulates pluripotency and histone H1 binding to chromatin. *Nature* **507,** 104–108 (2014).
60. Andrade, F. *et al.* Autocitrullination of human peptidyl arginine deiminase type 4 regulates protein citrullination during cell activation. *Arthritis & Rheumatism* **62,** 1630–1640 (2010).
61. Kearney, P. L. *et al.* Kinetic characterization of protein arginine deiminase 4: A transcriptional corepressor implicated in the onset and progression of rheumatoid arthritis. *Biochemistry* **44,** 10570–10582 (2005).
62. Keeling, P. J. & Palmer, J. D. Horizontal gene transfer in eukaryotic evolution. *Nat. Rev. Genet.* **9,** 605–618 (2008).
63. Husnik, F. & McCutcheon, J. P. Functional horizontal gene transfer from bacteria to eukaryotes. *Nat. Rev. Microbiol.* **16,** 67–79 (2018).
64. Moran, N. A. & Jarvik, T. Lateral Transfer of Genes from Fungi Underlies Carotenoid Production in Aphids. *Science* **328,** 624–627 (2010).
65. Chou, S. *et al.* Transferred interbacterial antagonism genes augment eukaryotic innate immune function. *Nature* **518,** 98–101 (2015).
66. Katz, L. A. Recent events dominate interdomain lateral gene transfers between prokaryotes and eukaryotes and, with the exception of endosymbiotic gene transfers, few ancient transfer events persist. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **370,** (2015).
67. Timmis, J. N., Ayliffe, M. A., Huang, C. Y. & Martin, W. F. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat. Rev. Genet.* **5,** 123–135 (2004).
68. Ku, C. *et al.* Endosymbiotic origin and differential loss of eukaryotic genes. *Nature* **524,** 427–432 (2015).
69. Hotopp, J. C. D. Grafting or pruning in the animal tree: lateral gene transfer and gene loss? *BMC Genomics* **19,** (2018).
70. Martin, W. F. Too Much Eukaryote LGT. *BioEssays* **39,** 1700115 (2017).
71. Martin, W. F. Eukaryote lateral gene transfer is Lamarckian. **2,** 754–754 (2018).
72. Leger, M. M., Eme, L., Stairs, C. W. & Roger, A. J. Demystifying Eukaryote Lateral Gene Transfer. *BioEssays* **40,** (2018).
73. Roger, A. J. Reply to 'Eukaryote lateral gene transfer is Lamarckian'. *Nature Ecology & Evolution 2018 2:5* **2,** 755–755 (2018).
74. Kondrashov, F. A., Koonin, E. V., Morgunov, I. G., Finogenova, T. V. & Kondrashova, M. N. Evolution of glyoxylate cycle enzymes in Metazoa: evidence of multiple horizontal transfer events and pseudogene formation. *Biol. Direct* **1,** 31 (2006).
75. Yue, J., Hu, X., Sun, H., Yang, Y. & Huang, J. Widespread impact of horizontal gene transfer on plant colonization of land. *Nat Commun* **3,** 1152 (2012).
76. Wybouw, N. *et al.* A gene horizontally transferred from bacteria protects arthropods from host plant cyanide poisoning. *eLife* **3,** (2014).
77. Schriek, S., Rueckert, C., Staiger, D., Pistorius, E. K. & Michel, K.-P. Bioinformatic evaluation of L-arginine catabolic pathways in 24 cyanobacteria and transcriptional

analysis of genes encoding enzymes of L-arginine catabolism in the cyanobacterium Synechocystis sp PCC 6803. *BMC Genomics* **8,** (2007).
78. Sanchez-Baracaldo, P., Ridgwell, A. & Raven, J. A. A Neoproterozoic Transition in the Marine Nitrogen Cycle. *Current Biology* **24,** 652–657 (2014).
79. Erwin, D. H. *et al.* The Cambrian Conundrum: Early Divergence and Later Ecological Success in the Early History of Animals. *Science* **334,** 1091–1097 (2011).
80. Yuan, X., Chen, Z., Xiao, S., Zhou, C. & Hua, H. An early Ediacaran assemblage of macroscopic and morphologically differentiated eukaryotes. *Nature* **470,** 390–393 (2011).
81. Falcao, A. M. *et al.* PAD2-Mediated Citrullination Contributes to Efficient Oligodendrocyte Differentiation and Myelination. *Cell Rep* **27,** 1090–+ (2019).
82. Kan, R. *et al.* Potential role for PADI-mediated histone citrullination in preimplantation development. *BMC Dev. Biol.* **12,** (2012).
83. Xu, Y. *et al.* Mutations in PADI6 Cause Female Infertility Characterized by Early Embryonic Arrest. *The American Journal of Human Genetics* **99,** 744–752 (2016).
84. Maddirevula, S. *et al.* The human knockout phenotype of PADI6 is female sterility caused by cleavage failure of their fertilized eggs. *Clin. Genet.* **91,** 344–345 (2017).
85. Qian, J. *et al.* Biallelic PADI6 variants linking infertility, miscarriages, and hydatidiform moles. *Eur. J. Hum. Genet.* **26,** 1007–1013 (2018).
86. Tietz, J. I. *et al.* A new genome-mining tool redefines the lasso peptide biosynthetic landscape. *Nat. Chem. Biol.* **13,** 470–+ (2017).
87. Wood, D. D., Moscarello, M. A., Bilbao, J. M. & O'Connors, P. Acute multiple sclerosis (marburg type) is associated with developmentally immature myelin basic protein. *Annals of Neurology* **40,** 18–24 (1996).
88. Spengler, J. *et al.* Release of Active Peptidyl Arginine Deiminases by Neutrophils Can Explain Production of Extracellular Citrullinated Autoantigens in Rheumatoid Arthritis Synovial Fluid. *Arthritis & Rheumatology* **67,** 3135–3145 (2015).
89. Zhou, Y. *et al.* Spontaneous Secretion of the Citrullination Enzyme PAD2 and Cell Surface Exposure of PAD4 by Neutrophils. *Front Immunol* **8,** 1200 (2017).
90. Stobernack, T. *et al.* A Secreted Bacterial Peptidylarginine Deiminase Can Neutralize Human Innate Immune Defenses. *mBio* **9,** 456 (2018).
91. Chory, E. J. *et al.* Nucleosome Turnover Regulates Histone Methylation Patterns over the Genome. *Molecular Cell* **73,** 61–72.e3 (2019).

Chapter 4: Activating PADI4 in Cells

4.1 Introduction

As reviewed in detail in the introduction, aberrant protein citrullination has been implicated in a diverse array of disease states[1]. This occurs through the deregulation of the PADI family of enzymes, which in most disease cases results in too much citrullination. This is particularly interesting due to the early promise shown by PADI inhibitors as therapeutic agents in disease models[2]. It has, nonetheless, proven very difficult to elucidate the mechanisms by which the PADIs are regulated physiologically and current understanding is limited beyond what is known *in vitro*. We therefore also lack mechanistic understanding of how the enzymes become deregulated in disease and how this contributes to disease phenotypes. The focus of this work is on understanding the physiological activation of one of the PADI paralogues: PADI4.

4.1.1 PADIs are regulated by calcium ions

PADIs are calcium-dependent enzymes. Although calcium is not directly involved in catalytic deimination, structural studies using PADI4 reveal that allosteric calcium ion binding drives an extensive conformational rearrangement from an inactive to active state of the enzyme (Figure 1.7 and 4.1)[3]. Five calcium ions bind such that the active site cleft becomes ordered and the nucleophilic active site cysteine moves ~5-10 Å into its catalytically competent conformation[3] (Figure 1.7).

Following the analysis of the binding site restructuring of PADI4 in *Arita et al.*, the crystal structures of the related paralogue PADI2 were solved in differentially calcium-bound states[3-5]. This revealed that sequential and cooperative calcium binding drives the conformational change: calcium binding in the PAD_M domain (calcium binding site Ca4) away from the active site also involves a residue in a loop from the catalytic PAD_C domain (D389), enabling the communication of allostery to the active site[5]. The

coordination of the first three calcium ions (sites Ca3-5) thereby aids binding of a final calcium ion near the catalytic site (Ca2), which is responsible for the restructuring of the active site (Figure 4.1). This also drives the movement of the nucleophilic active site cysteine. In the inactive conformation of PADI2 a 'gatekeeper' arginine residue acts as an intramolecular pseudosubstrate and shields the active site, but moves out of the active site in the active enzyme conformation and is replaced by the nucleophilic cysteine[5] (Figure 1.7). Both the calcium-switch region of the enzyme (Ca3-5) and gatekeeper arginine observed in PADI2 are also conserved in all active PADI paralogues and, although this has not been demonstrated in crystal structures of the other members, is likely to act as a switch for analogous regulation in the other paralogues of the PADI enzyme family– including PADI4.

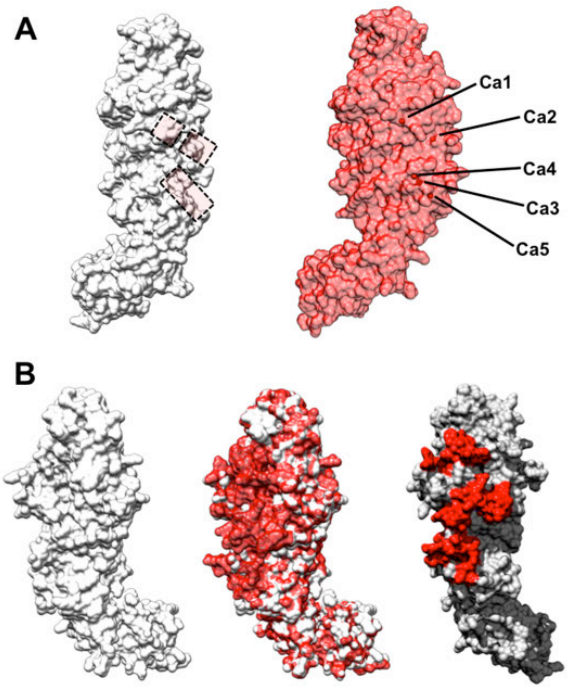


Figure 4.1: X-ray crystal structures of PADI4 with and without calcium ion binding. **A: Left panel:** X-ray crystal structure of PADI4 in the absence of Ca^{2+} binding shown coloured in white with protein surface displayed. The location of Ca binding sites 1 and 2 and Calcium binding site 3-5 are outlined in dashed transparent boxes. **Right panel:** X-ray crystal structure of PADI4 in the presence of 10mM $CaCl_2$ bound to 5 Ca ions (1wd9) shown coloured in red with protein surface displayed. The front face (right hand of the protein as shown) where the active site is located is now ordered. **B:** Three panels of a front view of PADI4. **Left panel:** Front view of PADI4 structure in the absence of Ca^{2+} shown in white

(1wd8). **Middle panel:** Calcium bound PADI4 structure 1wd9 shown in red was superimposed on the previous 1wd8 structure coloured in white using MatchMaker in Chimera where best aligning chains using the Needleman-Wunsch algorithm are iteratively matched by pruning long atom pairs until no pair exceeds 2.0 Å. **Right panel:** Calcium bound PADI4 structure is shown in gray. Regions, which were disordered (and therefore absent from 1wd8) in the calcium-unbound structure, that have become ordered upon calcium binding in 1wd9 have been coloured in red. Figures were prepared in Chimera using PDB structures 1wd8 and 1wd9 which were deposited by *Arita et al*.[3].

4.1.2 Comparing the calcium activation *in vitro* and in cells

The conformational switch described above, observed in PADI2, only occurs in the presence of a high concentration of calcium ions[5]. Concentrations of 10 mM $CaCl_2$ preserve the active form in crystallographic analysis[3]. Calcium switch regions are disordered and do not coordinate calcium ions at 50μM or 100μM, only becoming ordered at 250μM[5]. Detailed biochemical studies on the *in vitro* calcium dependence of the PADI enzymes have also been undertaken[6-9] and the calcium sensitivity for catalysis was found to be similar to the crystallographic studies (~250 μM Ca^{2+}). Using the citrullination lysate assay, I obtain similar results (Figure 4.2A). The half maximal calcium concentration for the artificial small peptide substrate N-benzoyl-l-arginine ethyl ester (BAEE) was found to be 0.5 mM, but for a real substrate histone 3, it was observed to be much higher at 3.3 mM[6,10].

In cells, however, PADI enzymes will not encounter such a high intracellular calcium concentration. Intracellular calcium is actively extruded from cells and maintained at a much lower level (~100nM)[11]. This is a requirement of all living cells not least because a high intracellular concentration of calcium ions would cause intracellular precipitation of DNA and RNA[11,12]. Studies measuring calcium in resting cells place the typical physiological intracellular calcium concentration at no greater than 100 nM[11]. Whilst intracellular calcium-signalling fluxes are widespread in cell biology, 1 μM would be a typical maximal physiological level, which is at least two orders of magnitude lower the requirement shown by PADIs *in vitro* (Figure 4.2A)[13].

One method to activate PADI4 in cells is to use a calcium ionophore (Figure 4.2B). In this instance, the elevated calcium concentration might be assumed (not unreasonably) to be responsible for the direct activation of the enzyme. Similarly, this reasoning ignores the scale of the calcium increase in cells. The calcium elevations caused by inflammatory stimuli and from ionophore in several studies have been measured, including in the same cell type where PADI4 is activated[13] and the maximum concentration of calcium inside the activated cells was measured to also be at 1 μM even after ionophore stimulation[13]. Subsequent studies have shown PADI4 can be activated in these exact cells under the same concentration of calcium ionophore and extracellular calcium in the buffer medium[14,15] and place the calcium requirement of PADI4 *in vitro* to be at least 100-fold higher. Other inflammatory stimuli (such as TNFα, or fMLF) can also cause intracellular calcium fluxes, but result in less extreme intracellular gradients than calcium ionophore and have been found to activate PADI4 robustly[16]. A particularly interesting inflammatory stimulus, lipopolysaccharides (LPS – isolated from the outer membrane of Gram-negative bacteria), is well characterized not to cause a spike in calcium levels during neutrophil stimulation, but still is observed to activate PADI4 strongly[17-19] (and personal communication, Prof Philip Cohen). A discrepancy clearly exists between the *in vitro* and physiological calcium requirement of PADI4; the activation by calcium of PADI4 *in vitro* appears to be insufficient to reconcile PADI4 activation in cells.

4.1.3 Previous efforts to understand PADI regulation

Given that the activation of PADI4 by calcium is unlikely to function similarly in the body, many people have hypothesized about alternative mechanisms of PADI regulation that might be relevant in vivo and significant efforts have been undertaken to progress this gap in the field.

One possibility is that the mobilization of free calcium ions inside cells, such as in chemokine receptor ligation or cellular differentiation, might achieve a local calcium concentration that is much higher than the average heightened

intracellular levels. This local concentration is likely to require an additional binding partner to mediate this. In particular, this would be required to explain how an intracellular elevation of calcium (or in the case of LPS, no elevation of cytosolic calcium, perhaps through a different mechanism) may be transduced to PADI4 in a specific and stringently regulated manner from such a diverse set of activating stimuli[1]. Efforts have been made to identify possible endogenous candidates, but these have so far remained elusive[20-22]. This is, nonetheless, an attractive hypothesis and there are indications this may be relevant for another paralogue, PADI3, in terminally differentiated keratinocytes as it may be mediated through the calcium binding protein S100A3[23]. A second important study showed that cross-reactive PADI3/PADI4 antibodies increase the calcium sensitivity of PADI4 *in vitro*, providing support for the possibility of an activating endogenous interacting partner[10]. Work described in Chapter 5, which was performed in parallel with the results of this chapter, in which Dr. Walport and I developed and characterized activating cyclic peptide molecules, provide further proof of principle that an interacting moiety can increase the calcium sensitivity of PADI4.

A second possibility is that an activating post-translational modification to the PADI sequence may prime the enzyme for activation. In light of this, previous studies assessed autocitrullination of the enzyme as well as the dimerization state of the enzyme, and despite demonstrating evidence that both phenomena occur, it was found that neither of these mechanisms explained PADI4 activation[20,24]. More promisingly, a paper analyzing a Rheumatoid Arthritis susceptible single nucleotide polymorphism (SNP) W620R in the phosphatase PTPN22, showed that the presence of the SNP disrupted the interaction of PTPN22 with PADI4 and resulted in increased citrullination[25]. Importantly, the paper indicated that PADI4 may be physiologically phosphorylated[25]. This finding was corroborated by data on PhosphoSite, which shows phosphorylated residues were observed on PADI4 from previous high throughput MS/MS proteomics searches (Figure 1.9B).

Along the same lines, a different study of PADI4 showed evidence that active and inactive species of PADI4 may migrate at different speeds by SDS PAGE and may indicate post-translational modification[21]. Another study showed that different protein kinase C (PKC) inhibitors caused various effects on PADI4 activation and inhibition[15]. *Neeli et al.* attribute the effects to specific inhibition of different PKC isoforms[15], but given these inhibitors have been shown elsewhere to be relatively unspecific[26-28], they may target many other kinases in the cell (binding at the ATP binding pocket). Nonetheless, the paper is interesting as it indicates that PADI4 activity may be regulated by kinase modulation. In addition, some PKC enzyme isoforms are themselves modulated directly by calcium and this therefore may explain at least in part some of the confusing correlations between calcium flux and calcium-dependent PADI activation observed in that paper[15]. Finally, another study revealed that different bicarbonate concentrations can affect both *in vitro* and cellular PADI4 activity. This is unlikely to have an effect on cellular calcium fluxes[29].

Given the lack of progress in identifying a candidate PADI activator in cells, the most common rationalization has been that since many substrates of the PADIs can be found in the extracellular space, particularly in synovial fluid (cytokines, collagen, fibronectin and fibrinogen, or even histones in post NETotic conditions), enzyme activity might be restricted exclusively to roles outside the cell where high concentrations of calcium are made available (extracellular calcium is as high as 1-2 mM in synovial fluid or blood plasma)[1,30,31]. This may go some way to explain the activity of PADIs in several innate immune contexts, such as after apoptosis, necrosis, or NETosis[17].

4.1.4 Re-evaluating the physiological activation mechanism of PADI4

It was the discovery of the role of PADI4 in the establishment of pluripotency[32] that prompted a reevaluation of the physiological activation of

the enzyme (Figure 4.2 and Figure 4.3). This newly discovered role of PADI4 appears to severely undermine the notion that PADI4 was only activated by obtaining access to extracellular calcium concentrations in dying or NETotic cells. In the context of the reprogramming of somatic cells to induced pluripotent stem (iPS) cells, mouse neural stem cells (which have no detectable PADI4 expression) begin to express PADI4 mRNA two days after the transduction of Yamanaka factors, and after eight days, are shown to possess activated PADI4, as interpreted by the detection of citrullinated histone 3[32]. The increase in citrullinated H3 occurs two days after being treated with two kinase inhibitors against GSK3β and MEK1/2 in the presence of Knock-Out Serum replacement (KSR), a treatment known as 2i (and referred to in this thesis as KSR2i)[32]. These iPS cells, with stably elevated levels of citrullinated histone 3 at the end of a course of reprogramming (after 14 days), are perfectly viable and can then be re-differentiated into any kind of cell. PADI4 can clearly be activated intracellularly and moreover in living cells (Figure 4.2C).

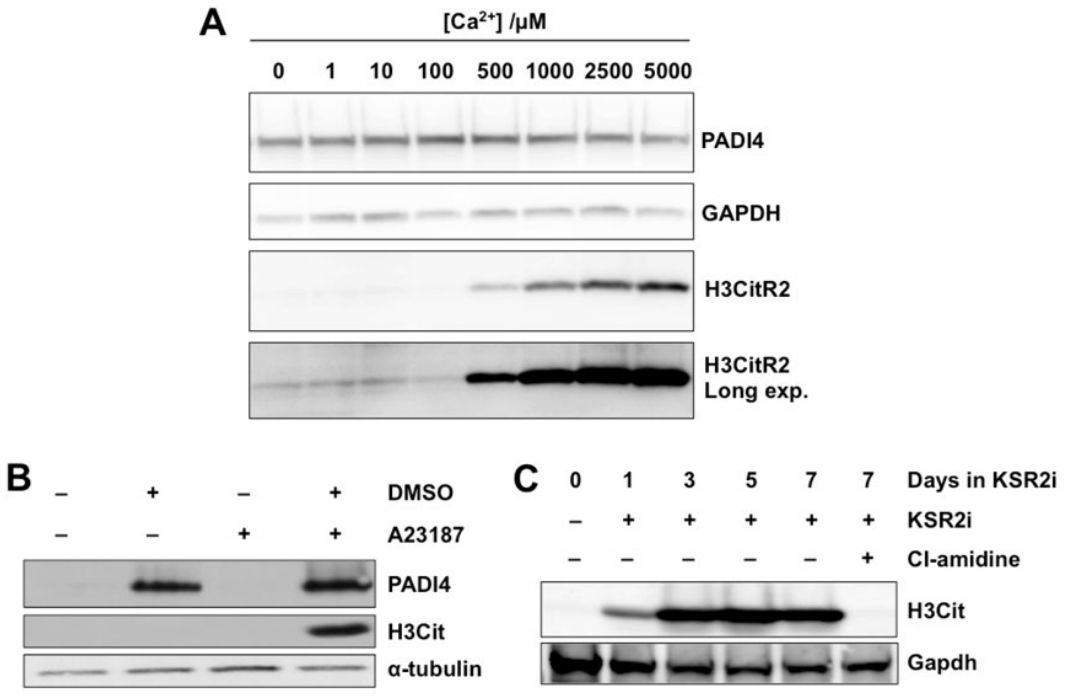


Figure 4.2: Rethinking the physiological activation of PADI4: comparing PADI4 activation in vitro, in HL60 cells and in mouse iPS cells. A: Immunoblot analysis of H3CitR2 of an in vitro citrullination lysate assay using mouse ES cells stably expressing

human PADI4. Serially diluted $CaCl_2$ was supplemented in each assay condition and incubated for 30 min at 37˚C. **B:** Immunoblot analysis of H3Cit and PADI4 of whole cell extracts from the HL-60 promyelocytic leukaemia cell line either cultured in RPMI with 10% serum (Lanes 1,3) or terminally differentiated into neutrophil-like cells by culturing in the presence of 1% DMSO (Lanes 2,4). After differentiation, PADI4 expression is induced and PADI4 protein can be detected. Cells were then treated with either vehicle (Lanes 1-2) or with the calcium ionophore A23187 (Lanes 3-4) in Locke's solution, before cell lysis and western blotting. **C:** Immunoblot analysis of whole cell lysates for a cellular reprogramming experiment, for H3Cit. Partially reprogrammed pre-iPS cells were cultured in KSR2i for 7 days (Lanes 1-5). Lane 6 shows cells treated in KSR2i for 7 days in the presence of Cl-amidine (200μM). Panel A is representative of n = 3. Dr Christophorou provided the blots for panels B and C, which reproduce established results from the literature[32,33].

This reevaluation seems important in addition for the well described role of PADIs in histone modification or transcriptional regulation[34,35]. PADI4 has been described in a variety of networks of transcription factors where chemical inhibition can reduce its interplay with other transcription factors and alter its function as detected by various chromatin immunoprecipitation (ChIP) analyses[35-37]. For this to be the case, resting activity of PADI4 must be able to take place in the nucleus of the cell. In these contexts calcium concentrations required *in vitro* are particularly implausible as the nucleus has been shown to be insulated from large or rapid cytosolic calcium increases[38].

4.1.5 Overarching aim

The primary aim of this chapter is to explore the possibility that there is more to the physiological regulation of the PADI enzymes than direct calcium binding or extracellular enzyme activity, and that some mediator helps to achieve the active and catalytic conformation of the enzyme inside cells either by increasing calcium sensitivity, enabling a high local calcium concentration, or in priming the protein for reactivity (these hypotheses are outlined in Figure 4.3). As such, one of the most critical questions in the citrullination field remains open – to elucidate how PADIs are activated physiologically and how this may become disrupted in the context of disease.

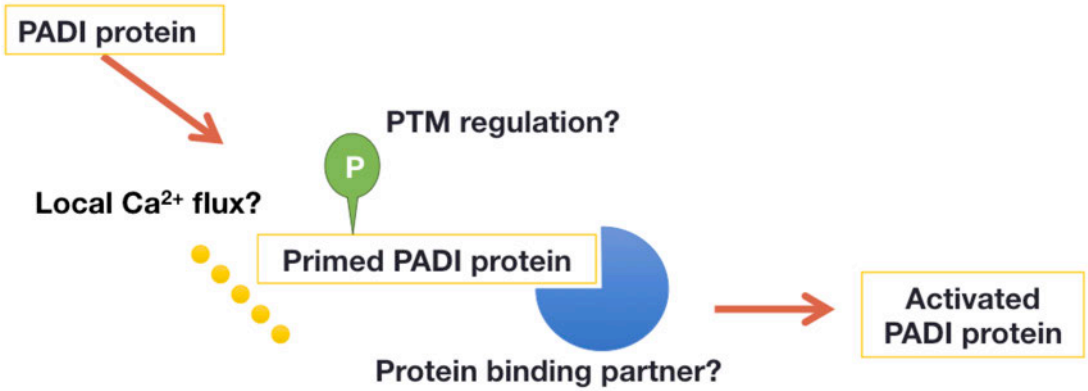


Figure 4.3: Schematic showing possible hypotheses concerning the physiological regulation of PADI protein in cells. Direct post-translational modification, local Ca^{2+} concentration or a protein-binding partner may enable PADI activation in vivo at a lower effective calcium concentration than that required *in vitro*.

4.1.6 Objectives:

- Establish cellular conditions for activating PADI4 tractable for detailed proteomic analysis.
- Elucidate signalling pathways upstream of cellular PADI4 activation

4.1.7 Overview of establishing tractable cellular systems for physiological PADI4 activation across innate immune and pluripotent cell contexts

As the lab was new at the start of my PhD, I began by setting up different cellular systems for PADI4 activation that may be tractable for proteomic analysis.

Firstly, it has been reported previously that PADI4 is activated in response to various pleiotropic inflammatory stimuli in primary human neutrophils and in terminally differentiated neutrophil-like cancer cells (HL-60 cancer cell line), responding as part of the innate immune system. In neutrophils, these stimuli are used to induce NETosis (or neutrophil extracellular trap formation)[17,39], which is the release of chromatin/DNA meshes that snare and kill infectious

agents[17,40,41]. Terminally differentiated HL60s are neutrophil-like cells but are poor at forming NETs, as only a low percentage are found to undergo NETosis, and show low antimicrobial activity[42,43]. NET forming stimuli, which have been reported in some cases to activate PADI4 in addition, include fMLF, LPS, IL-8, calcium ionophores (ionomycin and A23187), PMA, and TNFα. Reports in the literature are contradictory over whether PADI4 is essential for NETosis, and the role of PADI4 or citrullination in NET formation has not been clearly elucidated (see introduction for detailed discussion)[44-46]. Our interest, however, is not in NET formation per se, but on the mechanisms of PADI4 activation. I therefore began by setting up these two neutrophil or neutrophil-like systems in the lab and exploring different conditions to activate PADI4. The inflammatory stimuli that were shown to activate PADI4 in Neeli and Radic[16] were used as the starting point for this work (LPS, TNFα, fMLF and IL-8).

In tandem, I designed initial experiments to test the suitability of the cellular reprogramming system used in *Christophorou et al.* for assessing PADI4 activation[32]. In this system, pre-iPS cells (cells which have been frozen at day 6 after transduction of the Yamanaka factors) are grown in serum or in KSR2i conditions. The endogenous Oct4 locus is additionally tagged with GFP in this cell line (neural stem cells tagged with Oct4-GFP, NSO4G) giving proxy readout of reprogramming efficiency. Culturing the cells in serum results in a very low efficiency of reprogramming concomitant with a low extent of PADI activation as detected by histone citrullination[32]. Culturing in KSR/2i increases the efficiency of reprogramming to approximately 5% in parallel with the observation of increased citrullination of histone 3. It is known that inhibition or knockdown of PADI4 decreases the efficiency of reprogramming[32]. However, neither the amounts of PADI4 nor the population of reprogramming cells can be controlled in this system. This would confound any differential MS/MS analysis of interacting proteins between inactive and active conditions.

Two effects would need to be carefully controlled to discover molecular events leading to the enzymatic activity switch of PADI4. Firstly, levels of PADI4 protein would need to be controlled across conditions and independent from the stimulus used to drive PADI4 enzymatic activation. Secondly, the cell type in the system would need to be controlled and similarly independent from the stimulus used to induce PADI4 enzymatic activity.

As a result of the pitfalls of the reprogramming cell system, I decided to investigate activation conditions in a related mouse cell model system that is related to PADI4 in the pluripotent context, but which was controlled according to the essential criteria outlined above. I made use of an E14 line of mouse embryonic cell (mES) lines that stably express human PADI4 (mES PADI4 stable) and tested whether it might be possible to activate PADI4 in this cellular system. I used mES cells expressing an empty vector as a negative control (mES control stable) to isolate activation only of the exogenous protein. Despite the clear deviations of this cell line from a physiological situation (using exogenous protein expression), this presents a carefully controlled model system that could overcome the intrinsic deficiencies of the reprogramming system for the purpose of uncovering cellular mechanisms of PADI4 activation by MS/MS (detailed discussion Section 4.2.4). Subsequent validation of candidates in physiological systems will therefore be important.

4.1.8 Detection of PADI activity

Activation in cells was assessed by Western blot for the detection of an increase in histone 3 citrullination using two commercial antibodies: a polyclonal to H3CitR(2,8,17), referred to as H3Cit, and a monoclonal to H3CitR2 that became available during the work, referred to as H3CitR2. The first antibody is the best-characterized (and most sensitive) reagent to detect an increase in citrullination from cells, but is not without reported issues including lot-to-lot variability[47]. I carefully assessed all antibody lots in known

activation conditions prior to use and only used two lots that behaved in a controlled way under an established PADI4 activation stimulus (calcium ionophore activation in terminally differentiated HL60 cells) (Figure 4.2B, Figure 4.4). A specific problem was that some cross reactivity was observed with unmodified histone 3 so a clear increase in signal is required to conclude that enzymatic activation of PADI4 had occurred. Given the monoclonal antibody was shown to behave similarly to the polyclonal in our established condition, this was also employed during the work, but there may be contexts where the two antibodies do not behave identically. A whole cell lysis procedure was used for blotting, such that no material was discarded and histones could be reliably extracted to a comparable extent between conditions. Normalization was made to total cellular protein content as detected spectroscopically and verified using a loading control. A second detection method was with an antibody to chemically modified citrulline referred to throughout as Mod-Cit[48,49]. This approach chemically modifies the transferred blot with 2,3-butanedione monoxime, antipyrine, and $FeCl_3$ in a strong acid solution. The antibody detects the chemically modified moiety independent of sequence context, but given the antibody is no longer commercially available, its use is limiting and was therefore restricted to confirm certain findings. A third method using immunofluorescence detection of citrullinated histone 3 by high-content microscopy was also established to enable more high-throughput analysis in the near future and confirm activation was occurring in fixed nuclei and not merely in whole cell lysates (discussed in Section 4.2.6).

4.2 Results

4.2.1 Granulocyte-like cells differentiated from the HL-60 cell line

The HL-60 cell line is a human acute myeloid leukemia cell line, resembling promyelocytic cells, which can be terminally differentiated *in vitro* into granulocyte-like (neutrophil-like) cells using all-trans-retinoic acid or the polar solvent dimethyl sulfoxide (DMSO)[50,51]. Differentiation induces expression of PADI4, which in a resting state is not observed to be active (Figure

4.2B)[16,33,52]{Wang:2004hh}. To induce PADI4 activity, Neeli and Radic used a variety of inflammatory stimuli including fMLF, LPS, IL-8, TNFα, H_2O_2, ionomycin and calcium ionophore A23187[16]. I attempted to replicate these findings and found that calcium ionophore robustly induced activation of PADI4 in differentiated HL60 cells after 15 min (Figure 4.4A, Lane 9). In contrast to the Neeli and Radic paper[16], however, I was unable to activate PADI4 in differentiated HL60s using TNFα, LPS and IL-8 (Figure 4.4A, Lanes 2-4)[16,53].

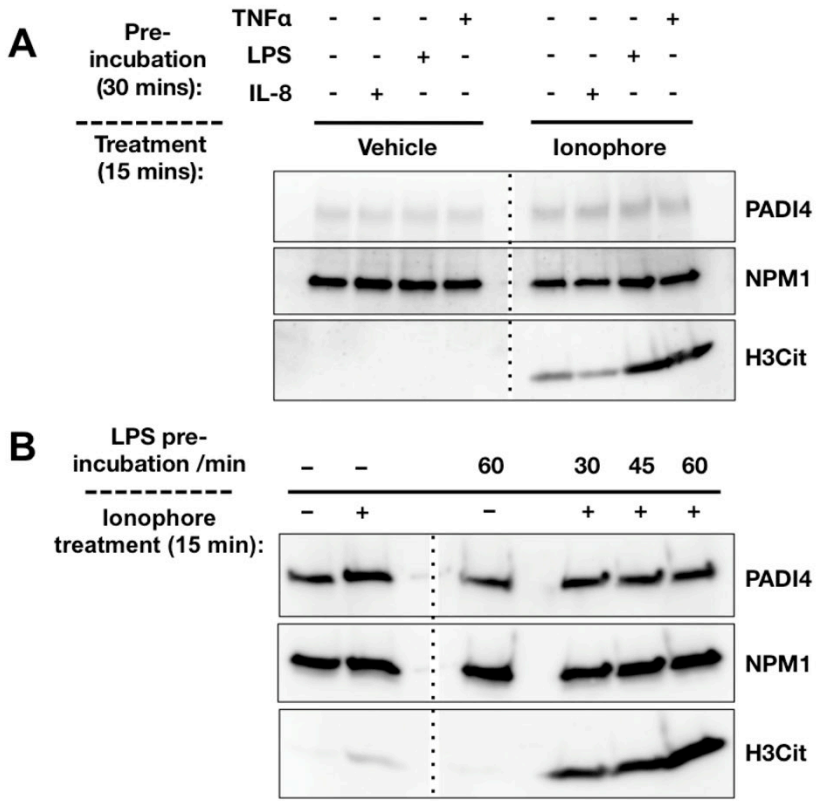


Figure 4.4: Activation of PADI4 by calcium ionophore in differentiated HL-60 cells is increased by priming with LPS pre-treatment. **A:** Immunoblot analysis of H3Cit of whole cell extracts from terminally differentiated HL-60 cells cultured in 1% DMSO for 5 days. Cells were treated with TNFα, LPS, IL-8 for 30 minutes and then treated with vehicle (Lanes 1-4) or A23187 (Lanes 5-8) for 15 minutes before lysis and blotting. **B:** Immunoblot analysis of H3Cit of whole cell extracts from terminally differentiated HL-60 cells (cultured in 1% DMSO for 5 days). Cells were treated with vehicle (Lanes 1-2) or LPS (Lanes 3-6) for up to 60 minutes before treatment with A23187 for 15 minutes (Lanes 2, 4-6). Data shown are representative of n = 3. Dotted line indicates the removal of irrelevant lanes from the same gel at the same exposure.

After speaking to members of the Rossi lab (MRC Centre for Inflammation Research, QMRI), I attempted to prime the cells for activation by pre-incubating with TNFα, LPS and IL-8 for 30 mins, as inflammatory stimuli are often used after a priming step in primary neutrophils, before attempting to activate PADI4 with calcium ionophore (personal communication, Dr David Dorward and Prof Adriano Rossi). Two out of three of these stimuli, LPS and TNFα, but not IL-8, increased the extent of activation of PADI4 by the ionophore A23187 (Figure 4.4A). From this preliminary result, I took the LPS priming treatment forward for validation given that it is known not to cause an intracellular calcium flux[18] (Figure 4.4B). Incubating LPS for longer (up to 60 minutes) increased the priming effect, causing a much greater extent of activation than without priming (Figure 4.4B, Lane 2 vs Lanes 4-6). It will be important to confirm that LPS was active by assessing rapid increase in levels of phospho-p38 after LPS stimulation (at ~30 min)[54]. These results are promising, as they indicate that there is more to activation of PADI4 than intracellular calcium flux provided by ionophore due to the priming effect by LPS, which is known not to mobilise intracellular calcium[18,55] (see hypothesis outlined in Figure 4.3).

4.2.2 Primary human neutrophils purified from peripheral blood

Differentiated HL60s have been shown to produce low numbers of NETs when compared to neutrophils[56], which may mean they are less relevant for the role of PADI4 activation. Given the difficulties encountered while attempting to reproduce the activation conditions in differentiated HL60s, it was decided to look at effects in primary human neutrophils purified from peripheral blood. Using a protocol developed in the Rossi lab (with initial instruction from Dr David Dorward, MRC Centre for Inflammation Research, Edinburgh), unstimulated primary human neutrophils were isolated from peripheral blood to greater than 95% purity, as measured by flow cytometric analysis and used for short term culture. Blood samples were drawn from healthy adult human volunteers and purified by the use of Dextran sedimentation and a discontinuous Percoll gradient (protocol in Chapter

2.2.4). The method results in the copurification of eosinophils, with distinctive orange granules under staining (Figure 4.5A, red arrow). Eosinophils, however, are typically at less than 1% abundance (but present at higher numbers in donors with hayfever).

Cells were subsequently cultured in LPS and fMLF, lysed rapidly and analysed by Western Blot (Figure 4.5B) with cell numbers that will be sufficient in the future for mass spectrometry analysis. Western blotting in primary human neutrophils is not straightforward; a rapid and stringent lysis procedure was employed to prevent activity of proteases. In contrast to the results obtained with differentiated HL-60s, fMLF and LPS activated PADI4 without the need for priming or treatment with Ionophore A23187 (Figure 4.5B). LPS treatment requires presence of 1% autologous serum, which is known to provide LPS binding protein as a cofactor for TLR-4 receptor stimulation[57].

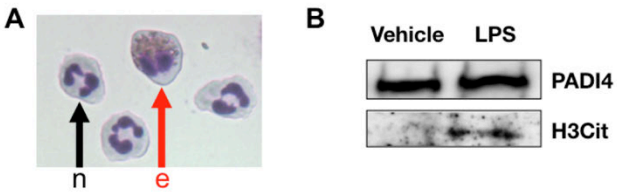


Figure 4.5: Purified primary human neutrophils treated with inflammatory stimuli show activation of PADI4. **A:** Cytospun image of purified granulocytes stained with DiffQuick. Neutrophils (black arrow, n) and eosinophils (red arrow, e) are purified, with neutrophils present at greater than 95% purity as determined by FACS. **B:** Immunoblot of whole cell extracts from primary human neutrophils elicited from healthy volunteers and purified by discontinuous Percoll gradient protocol (Chapter 2.2.4). After purification, cells were then cultured with fMLF (100 ng/mL) (Lane 1), vehicle (Lane 2), or LPS (1 µg/mL) in the presence of 1% autologous serum (Lane 3). Data shown are representative of n = 2.

Given the requirement for blood donation from healthy volunteers, the numbers of cells obtained are limited but it is notable that numbers of cells would nonetheless be sufficient for analysis by MS/MS. LPS treatment of purified human neutrophils represents an ideal physiological activating stimulus of PADI4. Firstly, it represents a physiologically relevant system,

involving endogenous levels of human PADI4. Secondly it is a rapid and physiologically relevant stimulus that models a real infectious situation. Thirdly, it is known not to cause an intracellular calcium flux[18] so is likely to reveal mechanisms that are not confounded by local calcium availability. It was considered that this cellular system would therefore be useful for validating candidates in a physiological system when required.

4.2.3 Mouse embryonic stem cells stably expressing human PADI4

I then looked to find conditions for activating human PADI4 in a context relevant to their role in pluripotency[32], but which was sufficiently controlled to be tractable for MS/MS analysis. To do this, I chose to make use of an E14 mouse embryonic stem (ES) cell line, which stably expresses human PADI4 (mES PADI4-stable). Pulldown of human PADI4 is achieved using an antibody to the endogenous human PADI4, but does not react with mouse PADI4 (personal communication, Dr Christophorou); no suitable antibody for pulling down mouse PADI4 was available commercially.

The piggyBac transposon system was used to stably overexpress human PADI4. The piggyBac system makes use of a piggyBac transposase, which recognizes inverted terminal repeat sequences (ITRs) located at either end of the sequence to-be-introduced, cuts it out, and introduces it across the genome at TTAA sites with high stability and efficiency. This has the advantage of incorporating the stably expressed gene across the genome. Levels of ectopic expression will therefore not be specifically dependent on a single cellular stimulus. Expression of human PADI4, in the absence of a global change in transcription, would therefore be constant and controlled. The endogenous PADI4 locus by contrast is under endogenous transcriptional control and therefore may change under different activating conditions without affecting the activation status of PADI4 protein per se, but it will not be under examination after affinity pull-down followed by MS/MS. Comparing the effects on mES cells that stably express an empty vector (mES Control-stable cells) reveals the extent of activation caused by the

exogenous human protein. This system therefore enables experiments that can deconvolute the effect of an activating stimulus on PADI4 protein activation from additional effects that may affect PADI4 protein levels. At the same time, however, there are clear downsides to using this ectopic system in which human PADI4 is stably expressed in mouse cells. For example, this approach could potentially reveal interactions and relationships between proteins or pathways that would not occur when proteins are expressed in cells from their cognate species at the endogenous level. As a result, it will be important to return to a physiological system for validation.

4.2.3.1 Treatment with calcium ionophore

To identify activating conditions of the exogenous hPADI4 protein, I first explored the effect of ionophore in both mES Control-stable and PADI4-stable cells. A low degree of background citrullination is observed in the untreated PADI4-stable cells over Control-stable cells (Figure 4.6A, comparing lane 1 and lane 7), implying some activity is observed in the resting state. This situation already contrasts with the observations in differentiated HL-60s (Figure 4.4). Mouse ES cells, despite having high and stable expression of exogenous hPADI4 were not activated by a 15-minute treatment of calcium ionophore (Figure 4.6A, lanes 7 and 8). By contrast, a robust increase in citrullination was observed at 6 hour and 24 hour time points after treatment with calcium ionophore in 10% serum (Figure 4.6A, lanes 10 and 12) suggesting activation of the exogenous human PADI4 was achieved. By comparing activation with the analogous treatment in control-stable mES cells, increases in citrullination are inferred to be a result of the exogenous human PADI4. It is interesting that after a 24-hour treatment with ionophore, a small activation of the endogenous mouse PADI4 can additionally be observed (Figure 4.6A, lane 6). The predominant effect, however, of activation of PADI4 enzyme in the 24-hour ionophore condition as detected by histone 3 citrullination is still due to the exogenously expressed hPADI4 (Figure 4.6A, lane 6 vs lane 12).

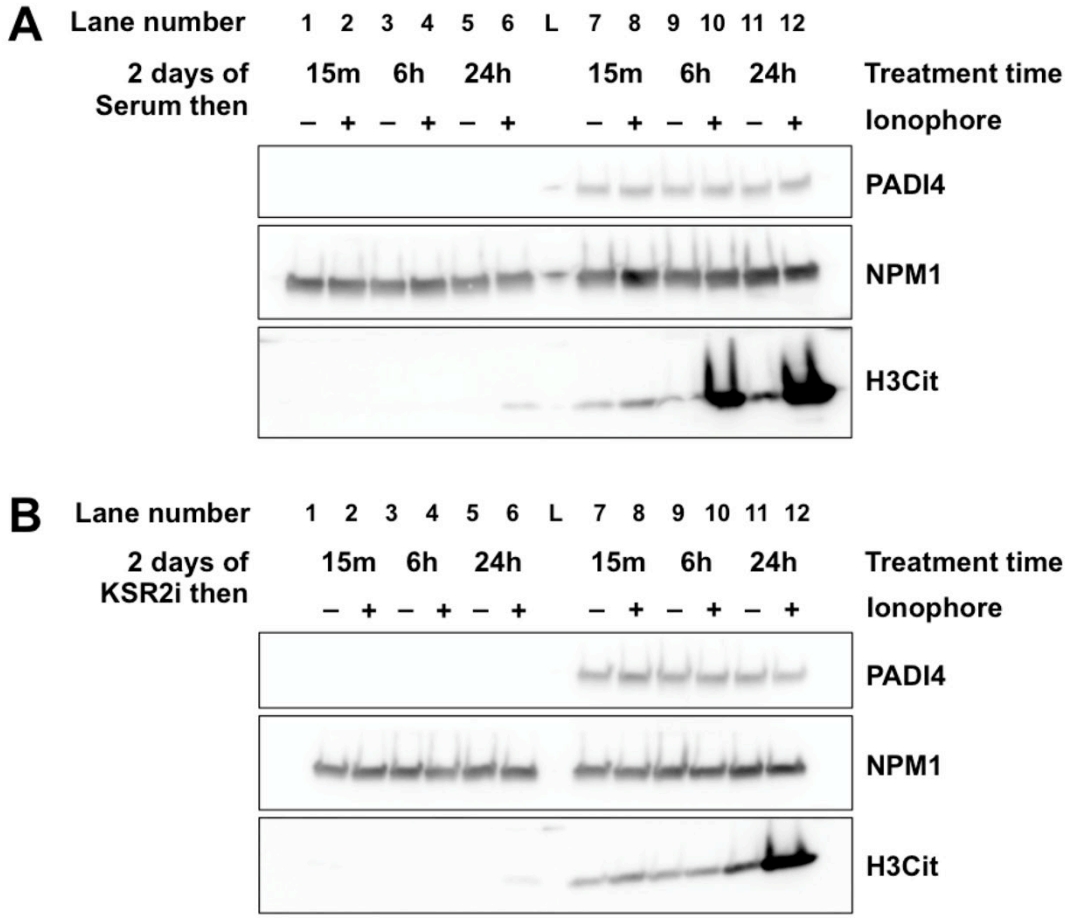


Figure 4.6: Activation of PADI4 in mES cells using calcium ionophore. Immunoblot analysis of H3Cit and human PADI4 of whole cell extracts of mES Control-stable (Lanes 1-6) mES PADI4-stable (Lanes 7-12) cells. Lanes labelled "L" contain protein molecular weight ladder. Mouse ES cells were cultured **A:** for 2 days in serum or **B:** for 2 days in KSR2i. Subsequently cells were treated with vehicle or A23187 for 15 minutes, 6 hours or 24 hours. Data shown are from a single experiment representative of n = 2.

The experiment was repeated in the presence of KSR2i (Figure 4.6B). KSR2i increases reprogramming efficiency when reprogramming somatic cells to iPS cells, but in mouse ES cells, this condition leads to the establishment of a ground state of pluripotency[58]. Control-stable and PADI4-stable mouse ES cells were treated in KSR/2i for two days, but 2 days in KSR2i did not activate PADI4 in contrast to reprogramming cells[32] (Figure 4.6B, lanes 1 and 7). As occurred in cells in serum, a 15-minute treatment of calcium ionophore did not activate mES PADI4-stable cells in KSR2i (Figure 4.6B, Lane 8). Surprisingly, the robust PADI4 activation after 6-hour ionophore which was

observed in 10% serum was completely abrogated in KSR2i (Figure 4.6B, lane 10 compared to Figure 4.6A, lane 10). By 24 hours, ionophore treatment did cause activation in KSR2i as it did for serum treated cells (Figure 4.6B, lane 12). A preliminary test to see if the activation of PADI4 in cells cultured in serum after 6 hours in ionophore (Figure 4.6A, lane 10) was dependent on transcription was performed by culturing cells in the presence of actinomycin D. This reduced the extent of PADI4 activation by calcium ionophore (Figure 4.7). Further work to try to deconvolute the effects of serum, KSR and 2i was then undertaken at different time points, but the effects were very complicated and will not be discussed in detail here, except to note that activation by ionophore at 15 minutes could be observed if the treatment was made in unsupplemented media, but was suppressed in the presence of serum or knock out serum replacement as observed in Figure 4.6A and B, lane 8).

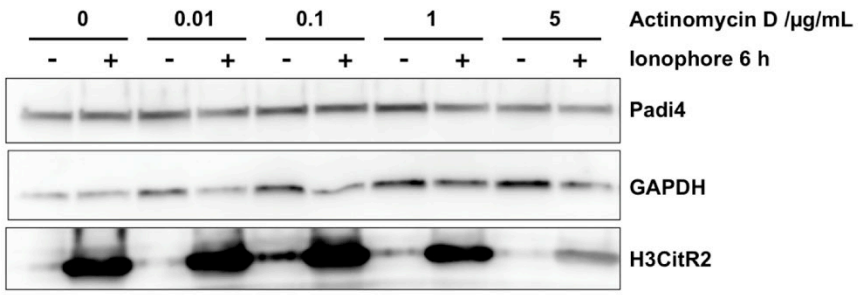


Figure 4.7: Activation of PADI4 in mES cells using calcium ionophore is reduced in the presence of actinomycin D. Immunoblot analysis of H3Cit and human PADI4 of whole cell extracts of mES PADI4-stable cells. Mouse ES cells were cultured for 2 days and treated with vehicle or 4µM calcium ionophore A23187 for 6 hours in the presence of an increasing concentration of actinomycin D. Data shown are representative of n = 2.

The complicated and interacting effects of calcium ionophore with serum or KSR on causing large changes in levels of citrullinated histone 3 were highly unexpected. For the purposes of choosing conditions for mass spectrometry, it was considered that further work would be useful to identify PADI4 activation conditions reliant on less complicated cellular stimuli. It would also be advantageous to identify activating stimuli that did not involve calcium ionophore. It was notable, however, that the unusual interactions of calcium

ionophore treatment and presence or absence of serum/ KSR would be consistent with complicated priming events occurring on PADI4 protein (as hypothesized in Figure 4.3).

4.2.3.2 Treatment with KSR2i

I then tested to see whether KSR2i, which activates PADI4 during the course of cellular reprogramming[58,59], might activate human PADI4 during a shorter treatment than 2 days. The levels of human PADI4 expression in mES PADI4 stable cells will not be affected by 2i treatment, unlike in the iPS cells where these effects cannot be separated. At 6 hours after switching into KSR/2i, PADI4 activity was robustly detected and was abrogated in the presence of the PADI inhibitor Cl-amidine (Figure 4.8A). Alpha-amanitin (α-amanitin) pretreatment (Figure 4.8B) and actinomycin D co-treatment (data not shown) were then tested with respect to KSR2i after a 6 hour treatment time point. Neither α-amanitin or actinomycin D abrograted PADI4 activation by KSR/2i in contrast to the effects on ionophore treatment. It will be important to confirm these effects occurred on transcription. However, this suggests that PADI4 activation by KSR2i at 6 hours is different in nature from the PADI4 activation caused by ionophore in these cells (Figure 4.8B) and implies PADI4 activation by KSR2i did not take place as a result of any transcriptional changes brought about in KSR2i, which might for example have resulted in the expression of an unknown activating protein factor. It was lastly checked against identical treatment on control stable cells where no activation was observed, suggesting activation was derived from exogenous human PADI4; this important control will need to be confirmed in further replicates (Figure 4.8C).

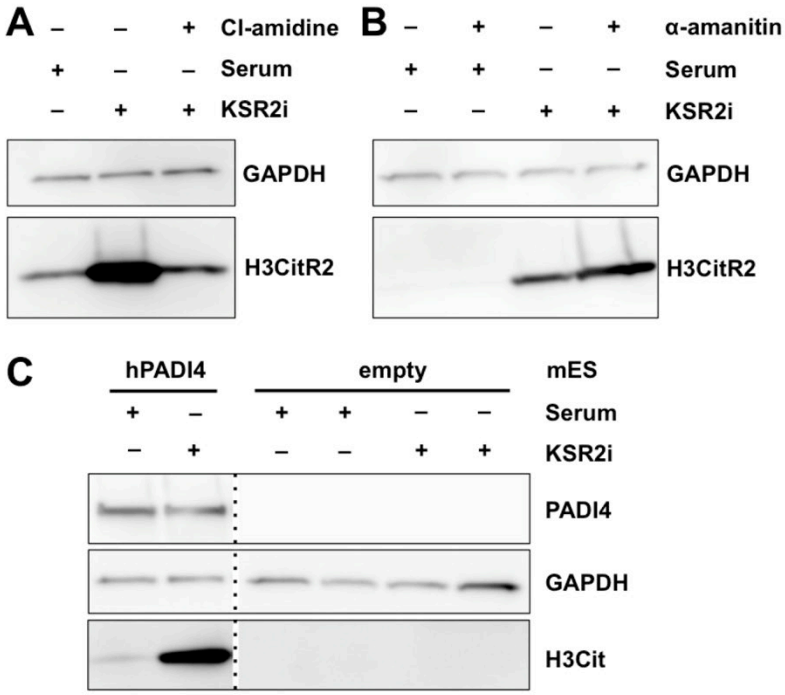


Figure 4.8: Activation of PADI4 in mES cells using KSR2i. **A:** Immunoblot analysis of H3CitR2 and Mod-Cit of whole cell extracts of mES PADI4-stable cells after six hour treatment with serum and vehicle (Lane 1), KSR 2i (Lane 2), and KSR 2i in the presence of 100μM Cl-amidine. 2i refers to the use of 1μM PD03250910 (inhibitor of MEK1/2) together with 3μM CHIR99021 (inhibitor of GSK3β)[58]. **B:** Immunoblot analysis of H3CitR2 of whole cell extracts of mES PADI4-stable cells after five hour treatment with vehicle (Lanes 1,3) or α-amanitin (Lanes 2, 4) followed by treatment for six hours of serum (Lanes 1-2) or KSR2i (Lanes 3-4) with α-amanitin added again for the second incubation. **C:** Immunoblot analysis of H3Cit and human PADI4 of whole cell extracts of mES PADI4-stable cells (Lanes 1-2) and mES Control-stable cells (Lanes 3-6) after 6 hour treatment in serum (Lanes 1, 3 and 4) or KSR with 2i (Lanes 2, 5 and 6). Dotted line indicates the removal of irrelevant lanes, from the same gel at the same exposure. In Panel A, data shown are representative of n = 3; in Panel B, data shown are representative of n = 2; in Panel C, data shown are from a single preliminary experiment.

4.2.3.3 Treatment using GSK3β inhibitors

Activation by KSR2i is likely to occur within the window of cell signalling, such as a signal transmitted by a kinase cascade. In the simplest instance this might be brought about by the action of the two kinase inhibitors that comprise 2i on those signalling pathways. To hone into the stimulus

responsible for activation of PADI4 within short-term KSR2i culture, I made efforts to deconvolute the different elements of KSR2i. In the first instance, I tested the effect of 2i in serum media, KSR media alone and 2i in KSR media (KSR2i) (Figure 4.9A). This indicated KSR and 2i act together to give full PADI4 activation at 6 hours, as the degree of activation by Serum2i is less than the degree of citrullination induced in KSR2i (Figure 4.9A). The activation effect was also confirmed using the Mod-Cit antibody (which recognizes protein citrullination independently from sequence context), showing that particularly the combined effect of KSR2i was robust in activating citrullination of a wide range of cellular substrates of different molecular weights (Figure 4.9A).

I then used the constituent inhibitors of 2i separately, culturing the cells in serum media. Treatment with the single kinase inhibitor CHIR99021 resulted in a small amount of rapid PADI4 activation (Figure 4.9B, C and D). This occurs as quickly as 15 minutes after inhibitor treatment and without changing media composition (in the presence of 10% serum throughout) (Figure 4.9B). Treatment with the MEK1/2 inhibitor alone did not result in any increase in citrullination by contrast, and the dual inhibition of 2i did not cause a greater activation than GSK3β inhibition alone (Figure 4.9B). This indicates that the effect can be ascribed to GSK3β inhibition (Figure 4.9B). In such a short time frame, it would be expected that effects of activation are occurring within the timeframe of cell signalling, but it will be useful to confirm this using α-amanitin and actinomycin D in future work. The effect of single GSK3β inhibition persisted at a 45 minute time-point (Figure 4.9 C and D). Structurally unrelated GSK3β inhibitors were then tested under the same conditions in mES PADI4-stable cells to validate the effect of CHIR99021 in acting on PADI4 was occurring via its published role of GSK3β inhibition and not through off-target action[26-28]. Short-term treatment using Li^+ or SB216763 after 45 minutes reproduced the effects of CHIR99021 suggesting inhibition of GSK3β is specifically responsible for PADI4 activation in this context (Figure 4.9C). This is consistent with the single digit nanomolar IC_{50} of

CHIR99021 for GSK3β inhibition in the literature (6.7 nM)[60] and studies which show it is a highly selective and specific small molecule inhibitor when used at the concentrations employed here[26-28].

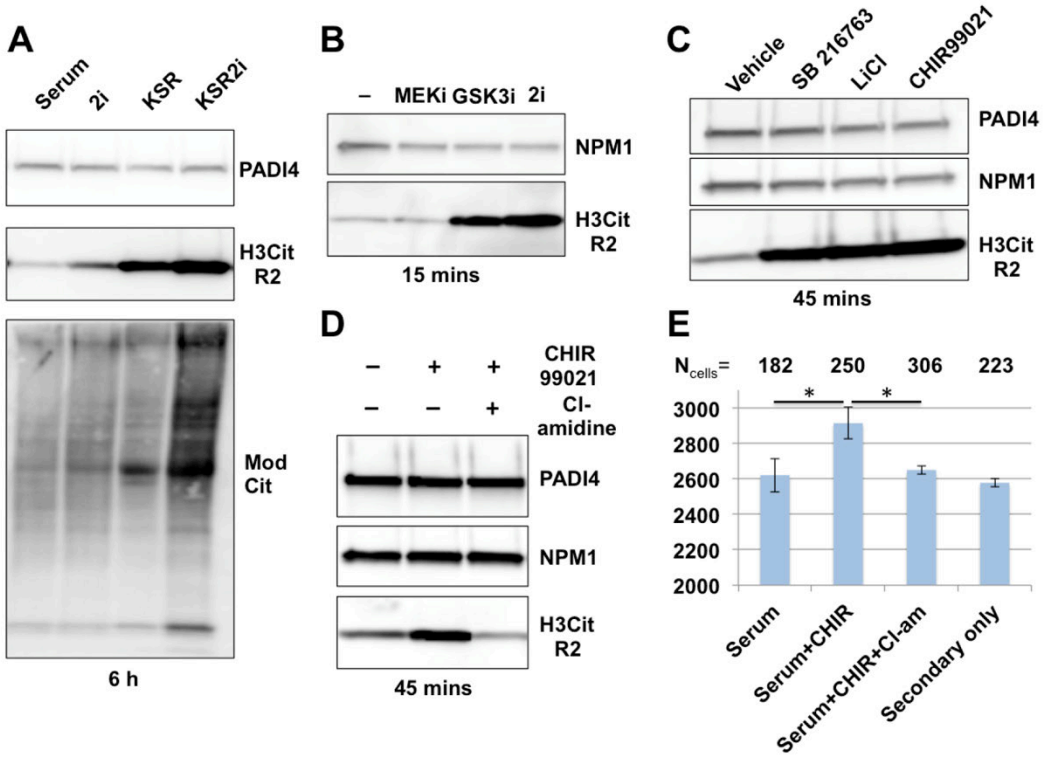


Figure 4.9: PADI4 is activated in mES cells after short-term GSK3β inhibition.

A: Immunoblot analysis of H3CitR2 and Mod-Cit of whole cell extracts of mES PADI4-stable cells after six hour treatment in serum (Lane 1), serum with 2i (Lane 2), KSR (Lane 3) and KSR with 2i (Lane 4). Dr Christophorou helped perform Mod-Cit analysis. Data are representative of n = 3; Mod Cit analysis was performed once. **B:** Immunoblot analysis of H3CitR2 in whole cell extracts of mES PADI4-stable cells cultured in serum for 2 days before 15 minute treatment in serum with vehicle (Lane 1), 1μM PD03250910 (Lane 2), 3μM CHIR99021 (Lane 3) or in 2i (Lane 4). Data are representative of n = 3. **C:** Immunoblot analysis of H3CitR2 in whole cells extracts from mES PADI4-stable cells cultured in serum for two days before treatment for 45 minutes with vehicle, 10μM SB216763, 10mM LiCl, or 3μM CHIR99021. Data are representative of n = 2. **D:** Immunoblot analysis of H3CitR2 in whole cells extracts from mES PADI4-stable cells after treatment for 2 hours in vehicle (Lane 1-2) or 100μM Cl-amidine (Lane 3) before 45 minutes with vehicle (Lane 1) or 3μM CHIR99021 (Lanes 2-3). The data shown are from the experiment performed in tandem with proteomic analysis from Chapter 6 and are representative of n = 3. **E:** PADI4-stable mES cells were grown sparsely in serum before treatment. Treatments comprised 15 min with vehicle (Bar 1), 10μM CHIR99021 (Bar 2 and 4) or pre-treatment in 100μM Cl-amidine before treatment with 10μM CHIR99021 for 15 min (Bar 3). Graph shows quantification of

H3CitR2 as determined by immunofluorescence with detection by high content microscopy (Chapter 2.4). The average nuclear intensity across two identically treated runs is shown. The number of cells analysed under high content microscopy is given above each bar and error bars show standard deviation and significance assessed using an unpaired student t test ($p<0.05$).

The effects of CHIR99021 in increasing H3Cit were then confirmed outside of Western blotting detection using immunofluorescence. To do this, I adapted a high-throughput immunofluorescence protocol from Dr Priya Hari and Dr Juan-Carlos Acosta[61] where fluorescence signal is captured using automated high content microscopy. Cells were treated with 10μM CHIR9901 for 15 mins or pretreated with Cl-amidine (200μM) before CHIR99021 treatment. Mouse ES cells form rounded 3D colonies in culture and do not grow easily in a single layer for high content microscopy detection so a protocol was established such that cells were seeded very sparsely before treatment (Chapter 2.4). At a 1:500 concentration of H3CitR2 antibody, signal was detected by immunofluorescence over background and over Cl-amidine treated cells (Figure 4.9E), reproducing the activation that is induced over background after 15 minute CHIR99021 treatment by Western blot (Figure 4.9B). In the future, this set-up should enable high-throughput analysis of PADI4 activation (see Discussion).

4.2.3.4 Note on PADI4 activation in mouse ES cells

At this point, a note should be made about the reproducibility observed on PADI4 activation as detected by histone 3 citrullination. The most robust effects are those presented, and those conditions using KSR 2i and the CHIR99021 were used for ~2 years in ~30 biological repeats across the various experiments described in this thesis. Unfortunately, during April 2018-June 2018 (at the end of my wet lab work), the behaviour of the mouse ES cells began to deviate from the results obtained previously. After the same treatment of KSR2i for 6 h or short term GSK3β inhibition, no increase in histone 3 citrullination was detected. During this period, I attempted to resolve this problem but could not provide a resolution in the time frame

available before stopping to finish the evolution work in Chapter 3 and to write the thesis.

Other labs had also noticed problems with cells cultured in 2i at the same approximate time (personal communication Dr Andrew Wood and various members of the ES cell culture room). During this period, I replaced all reagents, attempted a biological repeat in a different cell culture room, and tested different cell confluency, but encountered the same problems. It was reported to me while I was writing that the CO_2 levels in the building were increased in the three months after I finished and had been altered towards the end of the course of my work in the lab, so CO_2 consistency could be critically important. It would appear to this author that it is a high priority for future work to re-derive the PADI4 cell line in early passage mouse ES cells to attempt to reproduce the activation condition. It is also possible that later cell passages of stably expressed human PADI4 may adapt in culture and restrain or suppress the background levels of citrullination and that this could affect attempts to activate PADI4. It is likely in addition to be important to test and control different bicarbonate concentrations in the media, especially in light of the paper suggesting regulation of PADI4 activity in cells by bicarbonate concentration[29], given the knowledge that CO_2 levels changed over the course of the experiments. I had already submitted MS/MS conditions by this stage for which the activation was shown to be successful by Western blot (Figure 4.9D). These problems will be therefore be something to address before attempting future work but did not affect the work described in the rest of this chapter or Chapter 6. Until this can be resolved, any conclusions taken from this chapter should reflect a degree of uncertainty as to whether these effects represent normal cellular physiology.

4.2.4 Exploring PADI4 activation upstream of GSK3β inhibition: Wnt signalling

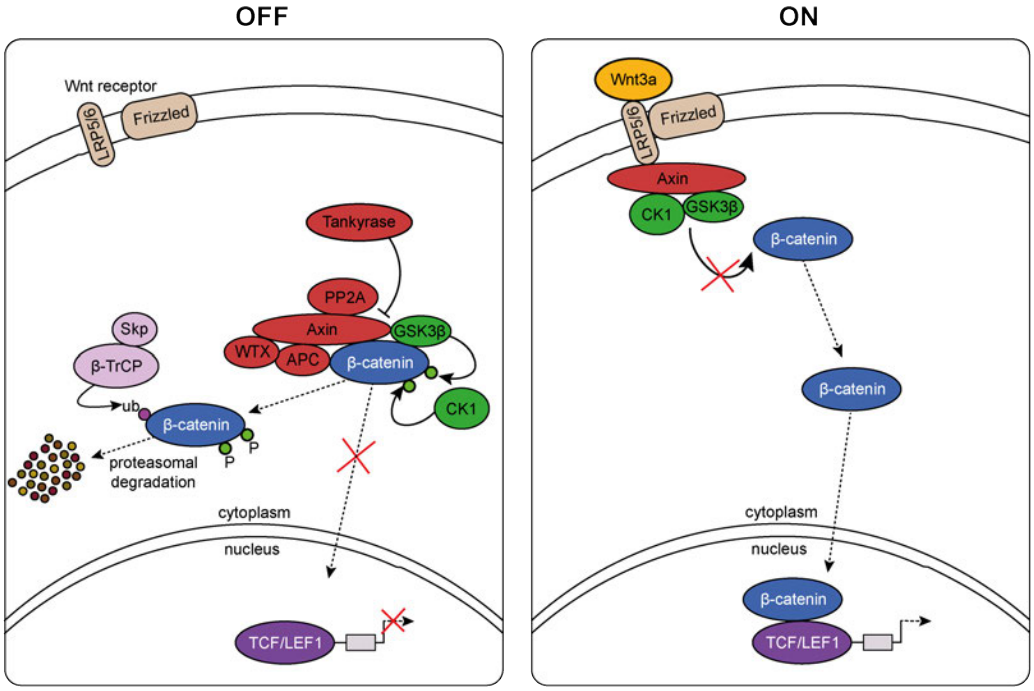


Figure 4.10: Simplified schematic of the canonical Wnt signalling pathway. The schematic is adapted from information in reviews of the Wnt signalling pathway[62,63].

To explore the signalling pathways activating PADI4 further, I then attempted to identify the upstream cause of GSK3β inhibition. GSK3β is a highly promiscuous kinase with a large number of possible cellular substrates. At least two different pools of GSK3β have been reported to act in the cell and act orthogonally in different signalling pathways[64]. Given that CHIR99021 is in part thought to mimic canonical Wnt signalling in the establishment of ground state pluripotency and in increasing the efficiency of reprogramming[58,65-70], I hypothesized that canonical Wnt signalling may be a good first candidate for identifying the signalling upstream of the kinase GSK3β that results in PADI4 activation.

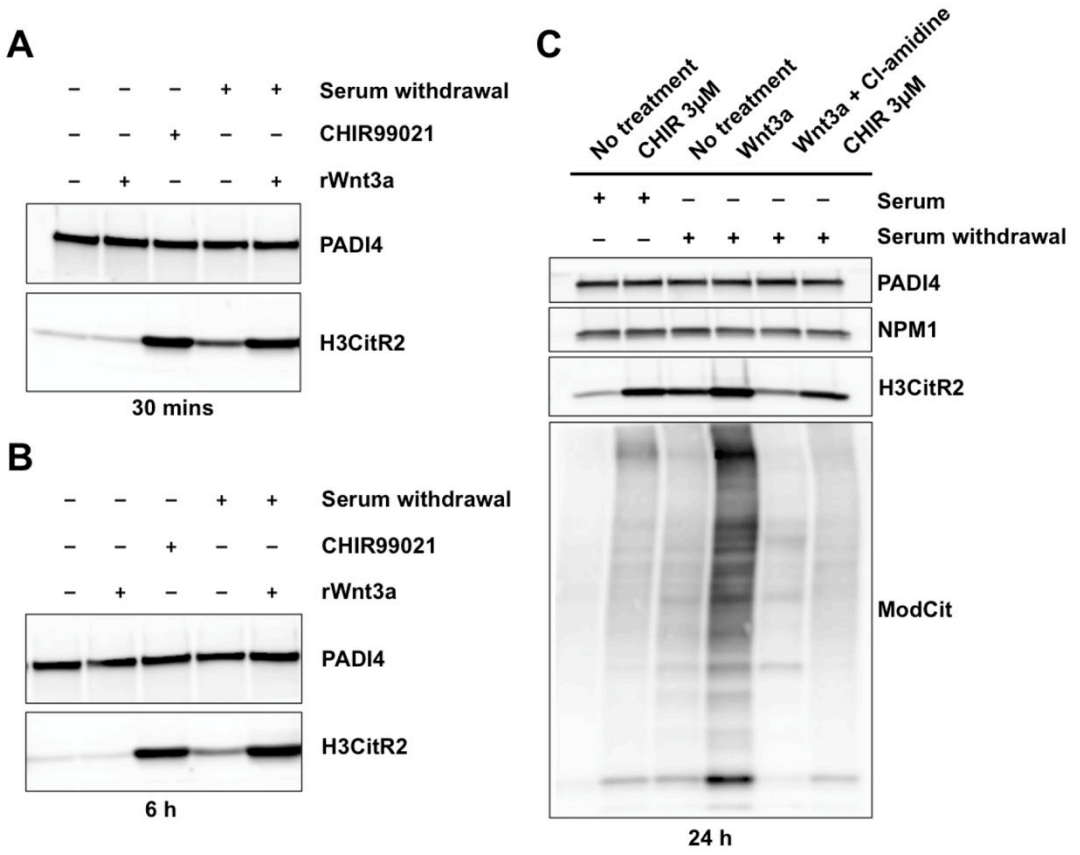


Figure 4.11: PADI4 is activated by treatment with Wnt3a. Immunoblot analysis of H3CitR2 or Mod-Cit of whole cells extracts from mES PADI4-stable cells cultured in serum for two days (**A-B:** Lanes 1-3; **C:** Lanes 3-6) or cultured in serum for 24 hours before serum withdrawal in KSR for 24 hours (**A-B:** Lanes 4-5; **C:** Lanes 3-6). Cells were then treated with vehicle, CHIR99021 3μM or recombinant Wnt3a 10ng/mL for 30 minutes (**A**), 6 hours (**B**), or 24 hours (**C**). In **C** Cl-amidine (100μM) added to cells 30 minutes prior to Wnt3a treatment (lane 5). Dr Christophorou performed Mod-Cit analysis. Data are representative of n = 2; Mod Cit analysis was performed once.

To test this, I used recombinant Wnt3a (10ng/mL) in place of CHIR99021 (3μM) on mES PADI4-stable cells. After 30 minutes, 6 hours and 24 hours of treatment, recombinant Wnt3a caused PADI4 activation, conditional on a 24 h serum withdrawal prior to treatment (Figure 4.11A, B and C). This was repeated using detection by anti-ModCit, showing that a broad range of substrates as well as histone 3 were also citrullinated after Wnt3a treatment (Figure 4.11C). Wnt3a activation after serum withdrawal strongly activates PADI4, doing so to a greater extent than CHIR99021 treatment as

determined by Mod-Cit detection. It is noticeable that this stronger extent of activation is not detected on H3CitR2 detection. We do not know whether histone 3 is the most important citrullinated substrate in this context and are merely using it as a useful proxy for PADI4 activation. These data therefore highlight the importance of validating PADI4 activity on other substrates in addition in future work. Optimizing an assay that can record activation of an alternative citrullinated substrate, in particular one that is less dynamically turned over than histone 3, such as using antibodies to citrullinated HP-1-gamma, NPM1, or hnRNPA1 may be particularly useful to support the work in this chapter.

Another small molecule Wnt agonist has been reported, BML284, which dose dependently induces β-catenin driven gene expression and is blocked in the presence of a dominant negative TCF4[71]. It is not known exactly how Wnt agonism is achieved directly by the small-molecule BML284, but BML284 does not inhibit GSK3β at these doses (IC_{50} > 60 μM)[71]. As further validation of the Wnt pathway, I therefore tested BML284 in PADI4-stable cells. BML284 activated PADI4 very robustly at 1μM and 10μM concentrations (Figure 4.12). No increase in H3Cit was observed on Control-stable cells.

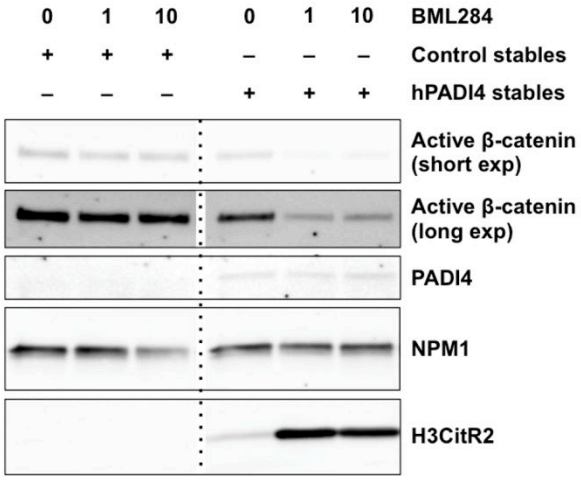


Figure 4.12: PADI4 is activated by BML284, a small molecule Wnt agonist. Immunoblot analysis of non-phosphorylated S45 of β-catenin, PADI4 and H3CitR2 of whole cells extracts from mES Control-stable cells (Lanes 1-3) or mES PADI4-stable cells (Lanes 4-6), cultured in serum for one day before 24 h in KSR media and treated with vehicle (Lanes 1, 4) or

BML284 (Lanes 2-3, 5-6) at final concentration of 1μM or 10μM. Dotted line indicates the removal of irrelevant lanes, from the same gel at the same exposure. Data are representative of n = 2.

These data would appear to place PADI4 activation in the Wnt signalling pathway, but suggest it occurs independently of GSK3β inhibition[71] (Figure 4.12). This would suggest that GSK3β inhibition therefore activates PADI4 for its effects on activating canonical Wnt signalling and that a downstream Wnt signalling effector goes on to activate PADI4. The other surprising observation was observed on the stability of β-catenin. Work was begun to explore these effects but will have to be addressed in the future.

4.2.5 Does PADI4 affect Wnt transcriptional output?

Taken together, these data suggest a role for canonical Wnt signalling (recombinant Wnt3a, BML284 and GSK3β inhibition) in activating PADI4 in this cellular system. Given the role for Wnt signalling in activating PADI4, I hypothesized that PADI4 may in turn play a role in modulating canonical Wnt signalling. Wnt signalling is thought to play a complicated role in the establishment of ground state pluripotency and in increasing the efficiency of reprogramming[62,63,72]. It would be particularly interesting if this may in principle be modulated by PADI4 activation, given that PADI4 inhibition decreases the efficiency of reprogramming. To test whether PADI4 may play a role in regulating the Wnt transcriptional output, I made use of a transcriptional reporter of Wnt signalling. Mouse ES cells stably expressing hPADI4 were transfected with a TOPflash vector, which contains binding sites for TCF/LEF that drive transcription of GFP, acting as a reporter of Wnt-mediated transcriptional activation. If Wnt signalling is activated, then the resulting stabilized β-catenin will drive GFP expression. A FOP vector is used as a control, which has mutations in the TCF/LEF binding site that render it inactive to Wnt driven transcription.

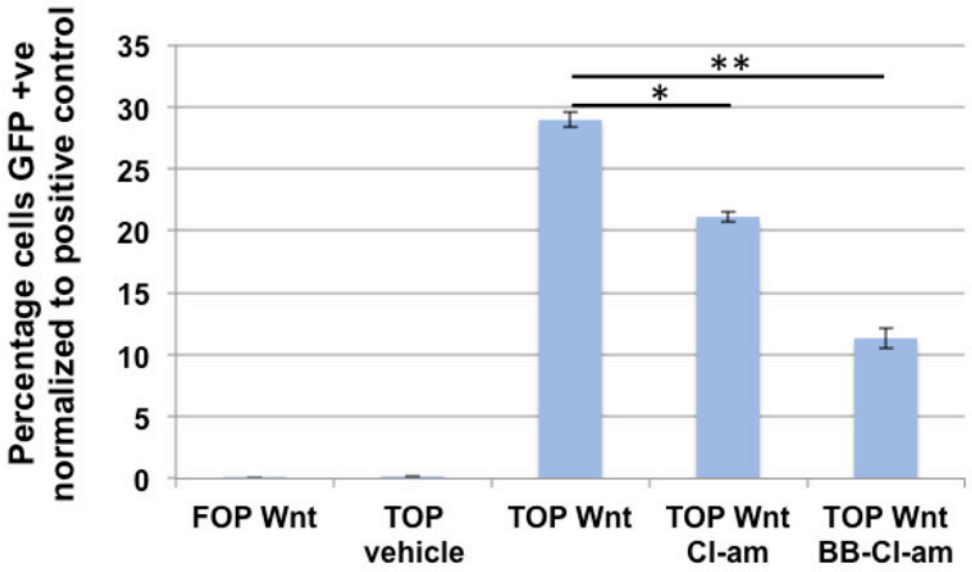


Figure 4.13: Inhibition of PADI4 reduces canonical Wnt signalling. PADI4-stable mES cells were cultured in KSR2i media for 24 hours, transfected with FOP-GFP or TOP-GFP and media replenished after 5 hours. Cells were then treated with vehicle, 100 ng/mL Wnt3a, 100 ng/mL Wnt3a and 100 μM Cl-amidine, or 100 ng/mL Wnt3a and 5 μM BB-Cl-amidine for 24 hours. GFP signal was detected by flow cytometry and was normalized to a GFP transfection control. Mean was taken of 3 biological replicates, error bars show standard deviation, and significance was assessed using a two tailed t-test (* $p < 0.05$, ** $p < 0.001$). Data shown are from a single experiment representative of n = 3

Treatment of mES PADI4-stable cells with Wnt3a caused an expected increase in Wnt-driven transcription as detected by the increase in GFP expression over untreated cells (Figure 4.13). This was not observed in the FOP control that was treated with Wnt3a (Figure 4.13). Inhibition of hPADI4 using Cl-amidine or BB-Cl-amidine (a more cell permeable and effective variant of Cl-amidine in cells)[73] caused a significant reduction in Wnt driven transcription (at least a two fold reduction by BB-Cl-amidine). These data show that Wnt driven transcription is reduced by PADI4 inhibition in this model system. This implies that PADI4 activity amplifies Wnt transcriptional output – such that Wnt signalling activates PADI4, which in turn amplifies the downstream gene expression. It will be very interesting to look at these effects using microarrays for Wnt target gene expression or by using RNA-sequencing to identify whether specific Wnt targets may be affected by

PADI4 activity. It will also be important to confirm these effects occur in a physiologically relevant system for Wnt signalling where PADI4 is also expressed such as in haematopoietic stem cells or in cancer contexts[58,62,63,65-70].

4.3 Discussion

In this chapter, I established stimuli for activating PADI4 in cells that are tractable for MS/MS analysis. In systems relevant both to innate immunity and pluripotency, this resulted in activation of PADI4 protein in cells where levels of PADI4 protein are carefully controlled and independent from the cellular stimulus. Mass spectrometry analysis of PADI4 activation using selected cellular conditions is then explored in Chapter 6 and this is discussed further there. In addition, work in this chapter looked at the signalling pathways upstream of PADI4 activation, which arose from the results obtained in this chapter. Excitingly, PADI4 was activated downstream of canonical Wnt signalling. This was confirmed using recombinant Wnt3a, the small molecule Wnt agonist BML284 and three distinct GSK3β inhibitors.

It is interesting that PADI4 can be activated by GSK3β inhibition, given that PKC inhibitors have previously been shown to activate or inhibit PADI4 under different conditions[15]. As these PKC inhibitors have been shown to inhibit at the ATP binding pocket, they target multiple other kinases and have been shown to target GSK3β[26-28]. It may be the case that the described activation of PADI4 by PKC inhibitors may be in part due to inhibition of GSK3β. A chemical genetic approach to explore the signalling pathways upstream of PADI4 was devised in conversation with Dr Greg Findlay and Prof Philip Cohen (MRC Protein and Phosphorylation Unit (PPU), University of Dundee). A small molecule inhibitor library to a large panel of kinases, which was a kind gift from Dr Greg Findlay and Prof Nathanael Gray, was made available to us. This was the primary reason for developing the high-throughput assay to detect PADI4 activation by immunofluorescence as previously I had relied on low-throughput detection by Western blotting to H3Cit. This high

throughput set-up will allow for these upstream signalling networks to be interrogated more carefully. In the first instance it will be interesting to test if other inhibitors in the library may mimic the activation caused by GSK3β inhibition. Secondly, using the library in combination with GSK3β inhibition or Wnt agonism (as a form of negative screen) is also likely to help dissect the pathway by which GSK3β inhibition or Wnt signalling activates PADI4. Kinase inhibitors that can disrupt PADI4 activation would be particularly interesting therapeutically.

In concert with the work looking at upstream signalling causing PADI4 activation, PADI4 inhibition was shown to reduce Wnt-driven transcription in the model ES cell system used in this chapter. It will be important to confirm that these effects on Wnt driven transcription occur in a more physiologically relevant context for Wnt signalling where PADI4 is also present[32,35,58,62,63,65-70,74,75]. That PADI4 inhibition can reduce Wnt-driven transcription is nonetheless an exciting finding as it could explain why inhibiting or knocking down PADI4 reduces reprogramming efficiency[32] – as the effect of this could be to concurrently dampen down Wnt signalling. Contexts where PADI4 is highly expressed may therefore be those that make use of enhanced Wnt signalling output. It is therefore particularly interesting that PADI4 is highly expressed in haematopoetic stem cells and upregulated in various cancer contexts[65,76,77], contexts which are well established to be reliant on increased Wnt signalling. Similarly this therefore provides a plausible hypothesis as to why PADI4 might be upregulated in certain cancer contexts. It will be interesting to know whether PADI4 activation may also play a role to redirect or tune Wnt signalling – gene expression microarray analysis or RNA sequencing experiments analogous to those in Figure 4.13 could test this hypothesis.

Further experiments will also help tease out mechanistic details by which increased PADI4 activity might enhance Wnt transcriptional activity, especially to look investigate effects on β-catenin stability and citrullination. It

is of note that while undertaking the experiments described in the chapter, another study identified PADI2 as having an opposing role in Wnt signalling[78]. In this study, a small molecule screen found a new role for the anti-parasitic drug NTX in inhibiting Wnt signalling. Following a chemoproteomic screen, the authors identified that NTX binds to and stabilizes PADI2. They subsequently showed that PADI2 citrullinates β-catenin directly and that this increases β-catenin turnover. In light of the results from the reprogramming system, possibly opposing roles of PADI2 and PADI4 might be anticipated.

This is interesting with respect to the data outlined here and also to a follow up experiment performed by Abigail Wilson, lab technician in the Christophorou lab. In the reprogramming system, two populations of cells were revealed (Figure 4.14). The reprogrammed population was enriched for PADI2, and the non-reprogrammed population enriched with PADI4. It is also notable that the non-reprogrammed population is enriched with PADI4 and could potentially show amplified Wnt signalling. This lends credence to the possibility that PADI2 and PADI4 might have opposing roles in Wnt signalling. If validated, this could imply a role for the non-reprogrammed cells as support cells that increase or drive reprogramming in the minority population. Given the role of Wnt in paracrine signalling, this is an attractive hypothesis.

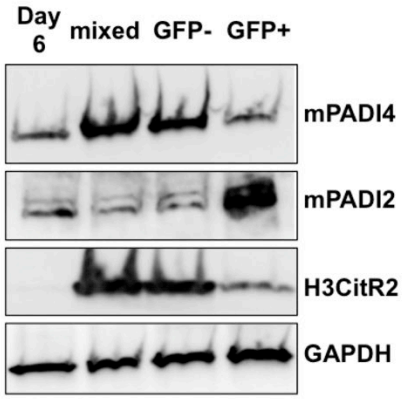


Figure 4.14: Contrasting expression of PADI4 and PADI2 in reprogramming cells. Pre-iPS cells (or day 6 cells) are neural stem cells, which have been transduced with Yamanaka

factors, are cultured in neural stem cell media for 4 days, then in mouse ES cell media supplemented with recombinant LIF for 2 days and finally frozen following *Christophorou et al.*[32,59]. **C:** Immunoblot analysis of mouse PADI4, mouse PADI2, and H3Cit of whole cell extracts of pre-iPS cells (day 6 cells, Lane 1). Pre-iPS cells were then also cultured for seven additional days of KSR2i treatment (Lanes 2-4). These cells (lanes 2-4) were then sorted GFP by FACS. Since the endogenous mouse Oct4 locus is tagged with GFP, after reprogramming, successfully induced pluripotent stem cells can be distinguished from the non-reprgrammed population by GFP expression. Immunoblot analysis was then performed on a mixed GFP and non-GFP population (Lane 2), a purified majority (~95%) non-GFP population (Lane 3), and on the purified minority (~5%) GFP-positive population (Lane 4). Experiments shown in this figure were performed by Abigail Wilson.

It will be important to consider the effects of both PADI2 and PADI4 in future experiments to explore in detail these effects. It would be interesting to repeat some of the experiments in this chapter using the analogous model system of PADI2-stable mouse ES cells such that the effects of Wnt signalling on PADI2 protein and the effect of PADI2 inhibition on Wnt signalling could be directly compared to the effects described for PADI4. Opposing effects for the two paralogues (PADI2 and PADI4) would be particularly important findings, especially in light of difficulties teasing apart their complementary roles in disease (discussed in Chapter 1). The data from this chapter therefore point to many exciting avenues for possible future work.

4.4 References for Chapter 4

1. Wang, S. & Wang, Y. Peptidylarginine deiminases in citrullination, gene regulation, health and pathogenesis. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1829,** 1126–1135 (2013).
2. Lewis, H. D. & Nacht, M. iPAD or PADi—'tablets' with therapeutic disease potential? *Current Opinion in Chemical Biology* **33,** 169–178 (2016).
3. Arita, K. *et al.* Structural basis for Ca2+-induced activation of human PAD4. *Nature Structural & Molecular Biology* **11,** 777–783 (2004).
4. Lewis, H. D. *et al.* Inhibition of PAD4 activity is sufficient to disrupt mouse and human NET formation. *Nat. Chem. Biol.* **11,** 189–191 (2015).
5. Slade, D. J. *et al.* Protein arginine deiminase 2 binds calcium in an ordered fashion: implications for inhibitor design. *ACS Chem. Biol.* **10,** 1043–1053 (2015).
6. Kearney, P. L. *et al.* Kinetic characterization of protein arginine deiminase 4: A transcriptional corepressor implicated in the onset and progression of rheumatoid arthritis. *Biochemistry* **44,** 10570–10582 (2005).
7. Knuckley, B. *et al.* Substrate specificity and kinetic studies of PADs 1, 3, and 4 identify potent and selective inhibitors of protein arginine deiminase 3. *Biochemistry* **49,** 4852–4863 (2010).
8. Darrah, E., Rosen, A., Giles, J. T. & Andrade, F. Peptidylarginine deiminase 2, 3 and 4 have distinct specificities against cellular substrates: novel insights into autoantigen selection in rheumatoid arthritis. *Ann Rheum Dis* **71,** 92–98 (2012).
9. Nakayama-Hamada, M. *et al.* Comparison of enzymatic properties between hPADI2 and hPADI4. *Biochemical and Biophysical Research Communications* **327,** 192–200 (2005).
10. Darrah, E. *et al.* Erosive Rheumatoid Arthritis Is Associated with Antibodies That Activate PAD4 by Increasing Calcium Sensitivity. *Science translational medicine* **5,** 186ra65–186ra65 (2013).
11. Williams, R. J. P. The evolution of calcium biochemistry. *Biochim. Biophys. Acta* **1763,** 1139–1146 (2006).
12. Eon-Duval, A., Gumbs, K. & Ellett, C. Precipitation of RNA impurities with high salt in a plasmid DNA purification process: Use of experimental design to determine reaction conditions. *Biotechnol. Bioeng.* **83,** 544–553 (2003).
13. Nüsse, O. *et al.* Store-operated Ca2+ influx and stimulation of exocytosis in HL-60 granulocytes. *J. Biol. Chem.* **272,** 28360–28367 (1997).
14. Zhou, Y. *et al.* Characterization of the Hypercitrullination Reaction in Human Neutrophils and Other Leukocytes. *Mediators of Inflammation* **2015,** 1–9 (2015).
15. Neeli, I. & Radic, M. Opposition between PKC isoforms regulates histone deimination and neutrophil extracellular chromatin release. *Front Immunol* **4,** 38 (2013).
16. Neeli, I., Khan, S. N. & Radic, M. Histone deimination as a response to inflammatory stimuli in neutrophils. *J Immunol* **180,** 1895–1902 (2008).
17. Brinkmann, V. *et al.* Neutrophil extracellular traps kill bacteria. *Science* **303,** 1532–1535 (2004).
18. Rodeberg, D. A. & Babcock, G. F. Role of calcium during lipopolysaccharide stimulation of neutrophils. *Infect. Immun.* **64,** 2812–2816 (1996).
19. Doerfler, M. E., Danner, R. L., Shelhamer, J. H. & Parrillo, J. E. Bacterial lipopolysaccharides prime human neutrophils for enhanced production of leukotriene B4. *J. Clin. Invest.* **83,** 970–977 (1989).
20. Slack, J. L., Larry E Jones, J., Bhatia, M. M. & Thompson, P. R. Autodeimination of Protein Arginine Deiminase 4 Alters Protein–Protein Interactions but Not Activity. *Biochemistry* **50,** 3997–4010 (2011).
21. Slack, J. L., Causey, C. P., Luo, Y. & Thompson, P. R. Development and Use of Clickable Activity Based Protein Profiling Agents for Protein Arginine Deiminase 4. *ACS Chem. Biol.* **6,** 466–476 (2011).
22. Zheng, L. *et al.* Calcium Regulates the Nuclear Localization of Protein Arginine Deiminase 2. *Biochemistry* **58,** 3042–3056 (2019).
23. Kizawa, K. *et al.* Specific Citrullination Causes Assembly of a Globular S100A3 Homotetramer: A Putative Ca2+ Modulatior Matures Human Hair Cuticle. *J. Biol.*

Chem. **283,** 5004–5013 (2008).
24. Liu, Y.-L., Chiang, Y.-H., Liu, G.-Y. & Hung, H.-C. Functional role of dimerization of human peptidylarginine deiminase 4 (PAD4). *PLoS ONE* **6,** e21314 (2011).
25. Chang, H. H., Dwivedi, N., Nicholas, A. P. & Ho, I. C. The W620 Polymorphism in PTPN22 Disrupts Its Interaction With Peptidylarginine Deiminase Type 4 and Enhances Citrullination and NETosis. *Arthritis & Rheumatology* **67,** 2323–2334 (2015).
26. Davies, S. P., Reddy, H., Caivano, M. & Cohen, P. Specificity and mechanism of action of some commonly used protein kinase inhibitors. *Biochem. J.* **351,** 95–105 (2000).
27. Mclauchlan, H., ELLIOTT, M. & Cohen, P. The specificities of protein kinase inhibitors: an update. *Biochem. J.* **371,** 199–204 (2003).
28. Bain, J. *et al.* The selectivity of protein kinase inhibitors: a further update. *Biochem. J.* **408,** 297–315 (2007).
29. Zhou, Y., Mittereder, N. & Sims, G. P. Perspective on Protein Arginine Deiminase Activity-Bicarbonate is a pH-independent regulator of citrullination. *Front Immunol* **9,** (2018).
30. Robertson, W. G., Marshall, R. W. & Bowers, G. N. Ionized Calcium in Body Fluids. *CRC Critical Reviews in Clinical Laboratory Sciences* **15,** 85–125 (1981).
31. György, B., Tóth, E., Tarcsa, E., Falus, A. & Buzás, E. I. Citrullination: A posttranslational modification in health and disease. *The International Journal of Biochemistry & Cell Biology* **38,** 1662–1677 (2006).
32. Christophorou, M. A. *et al.* Citrullination regulates pluripotency and histone H1 binding to chromatin. *Nature* **507,** 104–108 (2014).
33. Wang, Y. *et al.* Human PAD4 regulates histone arginine methylation levels via demethylimination. *Science* **306,** 279–283 (2004).
34. Cuthbert, G. L. *et al.* Histone Deimination Antagonizes Arginine Methylation. *Cell* **118,** 545–553 (2004).
35. Yuzhalin, A. E. Citrullination in Cancer. *Cancer Res.* **79,** 1274–1284 (2019).
36. Song, G. *et al.* A novel PAD4/SOX4/PU.1 signaling pathway is involved in the committed differentiation of acute promyelocytic leukemia cells into granulocytic cells. *Oncotarget* **7,** 3144–3157 (2015).
37. Ghari, F. *et al.* Citrullination-acetylation interplay guides E2F-1 activity during the inflammatory response. *Science Advances* **2,** e1501257 (2016).
38. Al-Mohanna, F. A., Caddy, K. W. T. & Bolsover, S. R. The nucleus is insulated from large cytosolic calcium ion changes. *Nature* **367,** 745–750 (1994).
39. Brinkmann, V. & Zychlinsky, A. Beneficial suicide: why neutrophils die to make NETs. *Nat. Rev. Microbiol.* **5,** 577–582 (2007).
40. Fuchs, T. A. *et al.* Novel cell death program leads to neutrophil extracellular traps. *J Cell Biol* **176,** 231–241 (2007).
41. Kenny, E. F. *et al.* Diverse stimuli engage different neutrophil extracellular trap pathways. *eLife* **6,** 178 (2017).
42. Papayannopoulos, V., Metzler, K. D., Hakkim, A. & Zychlinsky, A. Neutrophil elastase and myeloperoxidase regulate the formation of neutrophil extracellular traps. *J Cell Biol* **191,** 677–691 (2010).
43. Yaseen, R. *et al.* Antimicrobial activity of HL-60 cells compared to primary blood-derived neutrophils against Staphylococcus aureus. *J Negat Results Biomed* **16,** 2 (2017).
44. Li, P. *et al.* PAD4 is essential for antibacterial innate immunity mediated by neutrophil extracellular traps. *J Exp Med* **207,** 1853–1862 (2010).
45. Hemmers, S., Teijaro, J. R., Arandjelovic, S. & Mowen, K. A. PAD4-mediated neutrophil extracellular trap formation is not required for immunity against influenza infection. *PLoS ONE* **6,** e22043 (2011).
46. Liu, Y. *et al.* Peptidylarginine deiminases 2 and 4 modulate innate and adaptive immune responses in TLR-7 dependent lupus. *JCI Insight* **3,** (2018).
47. Neeli, I. & Radic, M. Current Challenges and Limitations in Antibody-Based Detection of Citrullinated Histones. *Front Immunol* **7,** 1532 (2016).
48. Senshu, T. *et al.* Detection of Deiminated Proteins in Rat Skin: Probing with a

Monospecific Antibody After Modification of Citrulline Residues. *Journal of Investigative Dermatology* **105,** 163–169 (1995).
49. Senshu, T., Akiyama, K., Ishigami, A. & Nomura, K. Studies on specificity of peptidylarginine deiminase reactions using an immunochemical probe that recognizes an enzymatically deiminated partial sequence of mouse keratin K1. *Journal of Dermatological Science* **21,** 113–126 (1999).
50. Collins, S. J., Ruscetti, F. W., Gallagher, R. E. & Gallo, R. C. Terminal differentiation of human promyelocytic leukemia cells induced by dimethyl sulfoxide and other polar compounds. *PNAS* **75,** 2458–2462 (1978).
51. Breitman, T. R., Selonick, S. E. & Collins, S. J. Induction of differentiation of the human promyelocytic leukemia cell line (HL-60) by retinoic acid. *PNAS* **77,** 2936–2940 (1980).
52. Hagiwara, T., Nakashima, K., Hirano, H., Senshu, T. & Yamada, M. Deimination of Arginine Residues in Nucleophosmin/B23 and Histones in HL-60 Granulocytes. *Biochemical and Biophysical Research Communications* **290,** 979–983 (2002).
53. Rohrbach, A. S., Arandjelovic, S. & Mowen, K. A. in *Protein deimination in human health and disease* (eds. Nicholas, A. P. & Bhattacharya, S. K.) 1–24 (Springer New York, 2014).
54. Malcolm, K. C. & Worthen, G. S. Lipopolysaccharide stimulates p38-dependent induction of antiviral genes in neutrophils independently of paracrine factors. *J. Biol. Chem.* **278,** 15693–15701 (2003).
55. Gupta, A. K., Giaglis, S., Hasler, P. & Hahn, S. Efficient neutrophil extracellular trap induction requires mobilization of both intracellular and extracellular calcium pools and is modulated by cyclosporine A. *PLoS ONE* **9,** e97088 (2014).
56. Remijsen, Q. *et al.* Dying for a cause: NETosis, mechanisms behind an antimicrobial cell death modality. *Cell Death & Differentiation* **18,** 581–588 (2011).
57. Muta, T. & Takeshige, K. Essential roles of CD14 and lipopolysaccharide-binding protein for activation of toll-like receptor (TLR)2 as well as TLR4 - Reconstitution of TLR2-and TLR4-activation by distinguishable ligands in LPS preparations. *European Journal of Biochemistry* **268,** 4580–4589 (2001).
58. Ying, Q.-L. *et al.* The ground state of embryonic stem cell self-renewal. *Nature* **453,** 519–523 (2008).
59. Theunissen, T. W. *et al.* Nanog Overcomes Reprogramming Barriers and Induces Pluripotency in Minimal Conditions. *Current Biology* **21,** 65–71 (2011).
60. Ring, D. B. *et al.* Selective glycogen synthase kinase 3 inhibitors potentiate insulin activation of glucose transport and utilization in vitro and in vivo. *Diabetes* **52,** 588–595 (2003).
61. Hari, P. & Acosta, J. C. Detecting the Senescence-Associated Secretory Phenotype (SASP) by High Content Microscopy Analysis. *Methods Mol. Biol.* **1534,** 99–109 (2017).
62. Clevers, H. & Nusse, R. Wnt/β-Catenin Signaling and Disease. *Cell* **149,** 1192–1205 (2012).
63. Nusse, R. & Clevers, H. Wnt/β-Catenin Signaling, Disease, and Emerging Therapeutic Modalities. *Cell* **169,** 985–999 (2017).
64. Doble, B. W. & Woodgett, J. R. GSK-3: tricks of the trade for a multi-tasking kinase. *J Cell Sci* **116,** 1175–1186 (2003).
65. Reya, T. *et al.* A role for Wnt signalling in self-renewal of haematopoietic stem cells. *Nature* **423,** 409–414 (2003).
66. Sato, N., Meijer, L., Skaltsounis, L., Greengard, P. & Brivanlou, A. H. Maintenance of pluripotency in human and mouse embryonic stem cells through activation of Wnt signaling by a pharmacological GSK-3-specific inhibitor. *Nat Med* **10,** 55–63 (2004).
67. Marson, A. *et al.* Wnt signaling promotes reprogramming of somatic cells to pluripotency. *Cell Stem Cell* **3,** 132–135 (2008).
68. Niwa, H. Wnt: what's needed to maintain pluripotency? *Nature Cell Biology* **13,** 1024–1026 (2011).
69. Wray, J. *et al.* Inhibition of glycogen synthase kinase-3 alleviates Tcf3 repression of the pluripotency network and increases embryonic stem cell resistance to differentiation. *Nature Cell Biology* **13,** 838–U246 (2011).

70. Yi, F. *et al.* Opposing effects of Tcf3 and Tcf1 control Wnt stimulation of embryonic stem cell self-renewal. *Nature Cell Biology* **13,** 762–770 (2011).
71. Liu, J. *et al.* A small-molecule agonist of the wnt signaling pathway. *Angew. Chem. Int. Ed. Engl.* **44,** 1987–1990 (2005).
72. Wray, J. & Hartmann, C. WNTing embryonic stem cells. *Trends Cell Biol.* **22,** 159–168 (2012).
73. Knight, J. S. *et al.* Peptidylarginine deiminase inhibition disrupts NET formation and protects against kidney, skin and vascular disease in lupus-prone MRL/lpr mice. *Ann Rheum Dis* **74,** 2199–2206 (2015).
74. Nakashima, K. *et al.* PAD4 regulates proliferation of multipotent haematopoietic cells by controlling c-myc expression. *Nat Commun* **4,** (2013).
75. Yuzhalin, A. E. *et al.* Colorectal cancer liver metastatic growth depends on PAD4-driven citrullination of the extracellular matrix. *Nat Commun* **9,** 4783 (2018).
76. Krivtsov, A. V. *et al.* Transformation from committed progenitor to leukaemia stem cell initiated by MLL–AF9. *Nature* **442,** 818–822 (2006).
77. Richter, J., Traver, D. & Willert, K. The role of Wnt signaling in hematopoietic stem cell development. *Critical Reviews in Biochemistry and Molecular Biology* **52,** 414–424 (2017).
78. Qu, Y. *et al.* Small molecule promotes beta-catenin citrullination and inhibits Wnt signaling in cancer. *Nat. Chem. Biol.* **14,** 94–+ (2018).

Chapter 5: Cyclic peptide reagents to target PADI4 for inhibition, activation and affinity purification

5.1.1 Introduction

The current toolkit for probing PADI4 activity is limited. PADI4 enzyme activity in cells is poorly understood and activation can be induced by pleiotropic stimuli. As there are no ways to activate the enzyme cleanly, this has hampered efforts to understand the biological functions of PADI4. Irreversible inhibitors have been developed[1,2], and a reversible inhibitor had been recently published at the start of my PhD that targets the inactive conformation of PADI4[3]. An enzyme activator would be particularly useful for the purpose of elucidating more precise biological roles for PADI4. It would also be useful to have additional tools capable of isolating the protein from cells, as well as other reagents for specifically visualizing PADI4.

This chapter describes the collaborative project that I initiated with Dr Louise Walport at the University of Tokyo to develop and characterise chemical biological tools for targeting PADI4. Dr Louise Walport was a colleague of mine during my Masters lab (Schofield lab, University of Oxford) and went to join the Suga lab (University of Tokyo) for a Marie Curie fellowship to study 2-oxoglutarate-dependent oxygenases[4] when I began my PhD. The Suga lab pioneered a method to screen a vast library of peptides for high binding affinity to a protein target[5]. Previously the RaPID method has been very successful in generating cyclic peptide displaying high affinity binding and potent inhibition of protein targets. We thought the RaPID method might be particularly appropriate for targeting PADI4 in order to generate potent reversible inhibitors. The RaPID system has not previously been used to find enzyme activators, but we considered that PADI4 may be an interesting test case for identifying PADI4 activators as well as potentially for finding reagents that can pull down the protein.

5.1.2 Challenges and progress in developing PADI inhibitors

One of the key challenges in discovering PADI inhibitors has been the development of high throughput assays for measuring protein citrullination. The colorimetric COLDER assay makes use of a Colour Developing Reagent (COLDER) (diacetyl monoxime and thiosemicarbazide in the presence of strong acids and Fe^{3+}) to convert citrulline into a visible dye, but the reliance of the assay on strong acids and toxic reagents has limited its use[6,7]. Glutamate dehydrogenase (GLDH) was used to monitor PADI activity through detection of ammonia release[8], but a significant problem is the need to deconvolute hits that target GLDH, which is not trivial given GLDH's own complex allosteric regulation. A competition assay involved incubating enzyme with target compounds in the presence of rhodamine-conjugated F-amidine, but relies on detection by gel electrophoresis (F-amidine and Cl-amidine are irreversible inhibitors discussed in the introduction, Section 1.9)[1,9]. This assay was used to identify various weak PADI inhibitors (such as streptomycin, chlortetracycline and minocycline as weak inhibitors). Additionally, an antibody based ELISA assay was developed but makes use of an antibody that is no longer commercially available[10]. Finally, a 264-member peptide library was synthesized (comprising variants of Ac-Y-X-F-amidine-cystamine) to adapt F-amidine for PADI paralogue selectivity and resulted in TDFA, which is an irreversible inhibitor that shows at least 15-fold preference for PADI4[11]. More recently, *Lewis et al*. used a fluorescence polarization binding assay in combination with GlaxoSmithKline's DNA-encoded small-molecule libraries in the presence and absence of calcium[3]. This identified a first compound GSK121 that is a weak inhibitor of PADI4[3]. After structure-activity relationship (SAR) development, this led to the optimized molecule GSK484, which targets the structure in the absence of Ca^{2+} and is a potent inhibitor at low concentrations of calcium (80nM), but shows five-fold reduced efficacy at higher concentrations of calcium[3,12]. *Tejeda et al*. subsequently used the COLDER assay to validate a set of compounds found through in silico binding approaches with selectivity for

PADI2, before SAR development led to an optimized lead[13]. This was tested for efficacy in a mouse model[13].

5.1.3 RaPID discovery system

The elegant Random nonstandard Peptides Integrated Discovery (RaPID) system was developed in the Suga lab (University of Tokyo)[5] and allows a huge library of peptides (containing ~10^{13} - 10^{15} peptides) to be screened for high binding affinity (low nanomolar K_D) to a target protein (Figure 5.1). This bypasses the need for a high throughput assay to measure protein citrullination as the first pass is high affinity binding, from which a number of candidates may be tested in more low throughput assays.

The RaPID system makes use firstly of flexizyme *in vitro* translation (FIT) using purified recombinant ribosomal components that can incorporate noncoded amino acids. Flexizymes are evolved ribozymes that are able to charge a variety of non-proteinogenic amino acids directly to the 3′-end of tRNA. In general, this is done simply by mixing the flexizyme, a matched acyl-donor, and tRNA in the presence of Mg^{2+} and incubating for 2 hours at 4°C. If a codon set is removed from the purified *in vitro* translation mix, then this native codon can be substituted for the flexizyme charged amino acid. In particular, FIT is typically used to replace the initiation codon (methionine) with an amino acid that spontaneously cyclises after translation. A non-proteinogenic chloroacetyl-amino acid (where the amino acid is typically Trp, Phe or Tyr) is charged to $tRNA_{fMet}$ and the machinery for methionine is removed from the *in vitro* translation mix. The produced peptide, with an N-terminal chloroacetyl amino acid then spontaneously cyclises post-translationally if a cysteine residue is present C-terminally in the peptide sequence. The cysteine attacks the chloroacetyl amino acid intramolecularly to form a macrocycle closed by a thioether bond. This reaction occurs spontaneously to form macrocycles of at least 20 amino acids long, independent of the peptide sequence composition. The constrained structure of the cyclic peptide means that it is held in a narrow distribution of

conformations, which aids the possibility of bioactive interactions. Macrocyclic peptides of this type are inspired by natural lantibiotics, which are similarly closed by a thioether linkage (such as that in lanthionine). Lantibiotics often also incorporate heavily post-translationally modified or otherwise non-proteinogenic amino acids (and may be the function of the cyanobacterial and other bacterial homologues identified in Chapter 3, Section 3.4). Using FIT, other amino acids can be similarly substituted as desired to replace another codon in the mixture enabling replacements such as non-proteinogenic amino acids, D-amino acids or backbone N-methylated amino acids. This additional variety, along with cyclisation, can confer on the polypeptide structural rigidity, very high target affinity, resistance to proteolysis, and occasionally membrane permeability.

Flexizyme *in vitro* translation was ingeniously combined with an mRNA-peptide display method originally developed in the Szostak lab[14,15] to comprise the full RaPID system, whereby cyclic peptide drug targets can be screened in a time frame of less than a week[5]. Overall, the RaPID system makes use of the following approximate general protocol. A vast DNA library is synthesized (coding for ~10^{15} peptides in a single screen – as much as eight orders of magnitude larger than most large pharmaceutical company small molecule screens). By comparison, the peptide library used to develop peptide versions of Cl-amidine such as TDFA made use of a library of 264 different peptides[11]. The vast DNA library is *in vitro* translated as described above to introduce the N-terminal chloroacetyl amino acid and a cysteine is always included C-terminally in the sequence, resulting in spontaneous peptide cyclisation. Crucially, this FIT step is also performed in the presence of the antibiotic puromycin, which is incorporated into the ribosome in the final step of translating messenger RNA for each peptide. Puromycin stalls the ribosome during translation such that each peptide is subsequently ligated to the mRNA that codes for its own sequence– thus acting as a barcode. Reverse transcription is then performed so that the barcodes are converted into mRNA-DNA hybrids. This vast library of barcoded cyclic

peptides is then presented to a target protein of interest (this process is referred to as a "round of selection"). Any peptide sequences that are retained (enriched in the selection) can therefore be amplified by polymerase chain reaction (PCR) and the process is repeated. After a number of rounds, certain peptides become heavily enriched and their barcodes can be amplified and sequenced. Any candidate peptides can then be synthesized by solid phase peptide synthesis (SPPS) and tested in downstream assays (Figure 5.1).

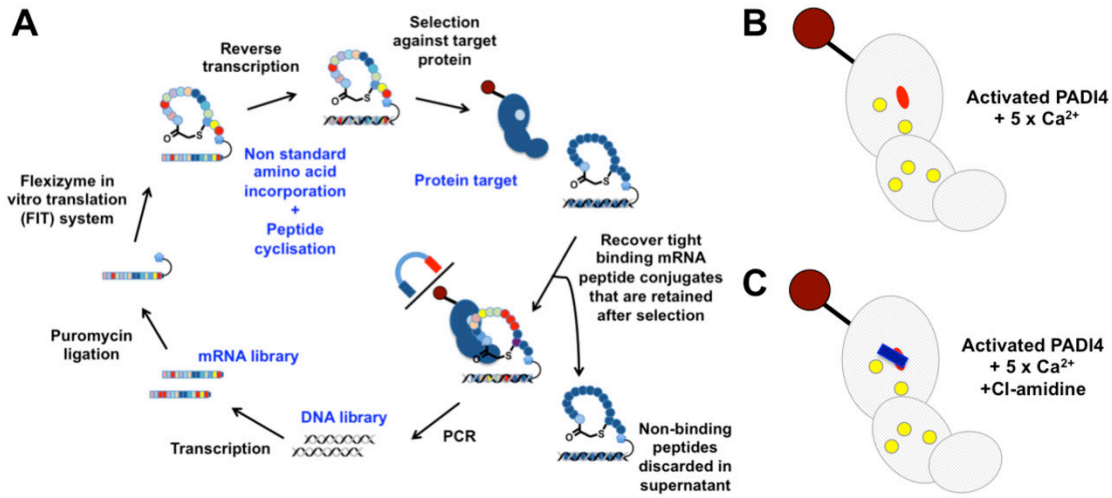


Figure 5.1: Schematic showing the RaPID system method and the design for PADI4 selections. **A:** RaPID overview. A DNA library (10^{13}) coding for short peptide sequences all containing an N-terminal methionine and a cysteine located C-terminally, is synthesized and transcribed into mRNA. The mRNA library is ligated to puromycin and *in vitro* translated using flexizyme charged tRNA to replace methionines with a chloroacetyl tyrosine amino acid. This generates a library of cyclic peptides each covalently ligated to the mRNA that codes for it. After reverse transcription, the library is washed against the target protein ('selection'). Tight binding peptides are retained and their sequences can be enriched by PCR. Further rounds of selection are performed which enriches sequences that bind with very high affinity. **B:** Schematic of the first selection to identify cyclic peptides that bind with high affinity to the active site of PADI4 in its active conformation. This was performed against biotinylated human PADI4 immobilised on streptavidin beads in the presence of 10 mM $CaCl_2$. See Figure 5.2A for crystal structure representation. **C:** Schematic of the second selection to identify cyclic peptides that bind allosterically to PADI4 protein trapped in the active conformation. This was performed against biotinylated human PADI4 immobilised on streptavidin beads in the presence of 10mM $CaCl_2$ but which was also incubated with Cl-amidine (an irreversible inhibitor that reacts covalently in the PADI4 active site to cysteine 645). See Figure 5.2B for crystal structure representation. **B, C:** Yellow circles represent

calcium ions, the red oval represents the enzyme active site, and the blue square represents the irreversible small molecule inhibitor Cl-amidine.

The RaPID system has been used for the discovery of many highly potent peptide inhibitors, typically with extremely high (low-to-sub nM) binding affinities (K_D) and IC_{50}s[5]. This has been used very successfully to develop potent inhibitors against a variety of protein targets. These include membrane proteins and extracellular enzymes including human α-amylase[16], drug transporters and receptor tyrosine kinases[17], but also inhibitors of some intracellular targets such as the AKT2 kinase[18], the ubiquitin ligase E6AP[19], the histone deacetylase SIRT2[20,21] and the lysine demethylase KDM4A[22]. Another peptide targeting a protein-protein interaction was discovered against Zaire Ebola virus protein 24 (VP24)[23], and allosteric inhibitors have also been found using this system[24]. Although in general the process reliably identifies molecules with very high binding affinity, the main focal problem for the broad application of cyclic peptides as therapeutic intracellular drug leads is that these macrocyclic inhibitors are only occasionally cell permeable and bioavailable.

It is not fully understood what elements of a peptide sequence, particularly in combination, may enable cell permeability, as there are a range of competing factors[25]. In some cases, cyclic peptides, such as cyclosporine A, are thought to be freely diffusible across cell membranes. In other examples, they are thought to hijack a cell's active uptake mechanisms: α-amanitin is thought to use an organic anion transporting polypeptide transporter[26], whereas cyanobacterial microcystins appear to hijack active uptake specific to mammalian liver cells (explaining their hepatoxicity)[27]. Other cell-penetrating peptides (or peptide transduction domains) appear to cross cell membranes directly. These short peptidic regions belong to diverse sequence classes ranging from arginine-rich polycationic sequences such as the HIV-1 TAT peptide (GRKKRRQRRRPQ)[28], through to amphipathic sequences such as the *Drosophila spp.* antennapedia peptide sequence

(RQIKIWFQNRRMKWKK)[29], and even to a rarer class of predominantly hydrophobic sequences[30-32].

The secondary structure of the peptide, such as the presence of alpha helical structure, is also likely to be involved in permeability, as is polarity, size and lipophilicity. Other peptidic molecules may traverse membranes by forming pores, through endocytosis mechanisms, such as macropinocytosis, or through hijacking the endoplasmic reticulum associated protein degradation (ERAD) machinery. Because of the multiple methods by which cyclic peptides might attain cellular access and although certain peptides can be engineered to ameliorate their intrinsic cell permeability, a general and predictive approach has not yet been elucidated and adopted. Multiple strategies can nonetheless be employed to improve intrinsic cell permeability including backbone N-methylation, single point substitutions, D-amino acid incorporation, myristoylation, tagging to a cell targeting peptide, or conjugating to a scaffold with intrinsically permeable properties[33]. In individual cases, it is difficult to predict in advance which may be the most valuable. An alternative and likely promising approach would be a cargo approach for therapeutic delivery[25,34].

5.1.4 Previous RaPID targeting of PADI4 in the Suga lab

As it happens, several years before the start of my PhD project, a C-terminal His-tagged-PADI4 had already been screened for inhibitors in the Suga lab against a newly designed library that had been created to optimize cell permeability (personal communication, Prof Hiroaki Suga and Dr Louise Walport). The selection was carried out without calcium in the buffers and with ethylenediaminetetraacetic acid (EDTA, a calcium ion chelator) present. Several 30-80 nM K_D binders were found as determined by surface plasmon resonance (SPR) in the absence of Ca^{2+} and in the presence of EDTA. However, *in vitro* IC_{50}s by the COLDER assay were only ~5 μM and only slightly more potent than Cl-amidine. COLDER activity assays were carried out after pre-incubation of PADI4, peptide and calcium for 15 minutes. The

most potent peptides all included an arginine residue; it was not tested whether this was converted to citrulline. A second-generation peptide was made where the arginine was replaced by Cl-amidine, but which did not increase the potency greatly. Activity assays were subsequently performed in cells, using western blot to detect citrullination, but inhibition was not observed using the peptides from the selection.

Dr Louise Walport and I hoped to try to improve on the results obtained previously and in particular to design selections that might result in a greater correlation between binding affinity and potency. It was therefore decided to subclone the PADI4 gene into a 6xHis-Avi-tagged expression vector and co-express the protein with the biotin ligase BirA. The 6xHis tag allows for high yield and efficient purification. The Avi tag is site-specifically biotinylated by BirA at a Lys residue contained within the tag, such that the protein can be coordinated to streptavidin-coated magnetic beads during the selection step. The strong affinity of biotin for streptavidin enables robust protein immobilization during the RaPID system process and in general is found to give better results (such as by allowing for more stringent washing steps). We also opted to design modified selection strategies to try to improve the correlation between binding affinity and potency of inhibition.

5.1.5 Objectives

- Design and perform cyclic peptide selections to target PADI4
- Screen candidate cyclic peptides targeting PADI4 for inhibition, activation and pulldown
- Characterise peptides for activity *in vitro* and in cells and optimise peptides for cellular permeability

5.2 Designing selections to target PADI4

Dr Louise Walport and I designed two modified selection strategies for the identification of PADI4-binding peptides. Firstly, it was decided to preincubate the immobilized beads bound to PADI4 protein in a high calcium ion

concentration (10 mM Ca^{2+}) before performing the selection (Figure 5.2A). The high calcium concentration would be maintained in all buffers during the selection steps. It was reasoned that this would force the PADI4 enzyme into an active conformation for the selection rounds, such that high affinity binding peptides might be found against the fully structured active site cleft[35]. It was hoped that selecting against the active enzyme conformation, provided that calcium did not interfere with the RaPID steps, would be likely to give an improved correlation between binding affinity and inhibition.

We then designed a second selection. PADI4 would be preincubated in calcium containing a saturating concentration of Cl-amidine prior to selection. Cl-amidine is an irreversible small molecule pan-PADI inhibitor that reacts covalently with the active site cysteine of PADI4 and therefore blocks access to the active site (Figure 5.2B). This inhibited form of PADI4, maintained in high calcium and trapped in an active conformation, would then be used in the rounds of selection. It was hoped therefore that this strategy might enrich for high affinity binding peptides that bind PADI4 allosterically. Within such a pool of peptides, we hoped that we might identify ones that might act to stabilize the active conformation of the enzyme and act as activators of the enzyme. This would in theory provide proof of concept that the allosteric binding of an interacting partner can activate PADI4, which has been hypothesized to occur within the cell.

Finally, candidate peptide sequences that bind strongly and independently of the calcium concentration are likely to be enriched in either selection. These would be potentially useful for developing into affinity reagents or a chemical biological probe for visualizing PADI4 by, for example, conjugating it to biotin or a fluorescent reporter.

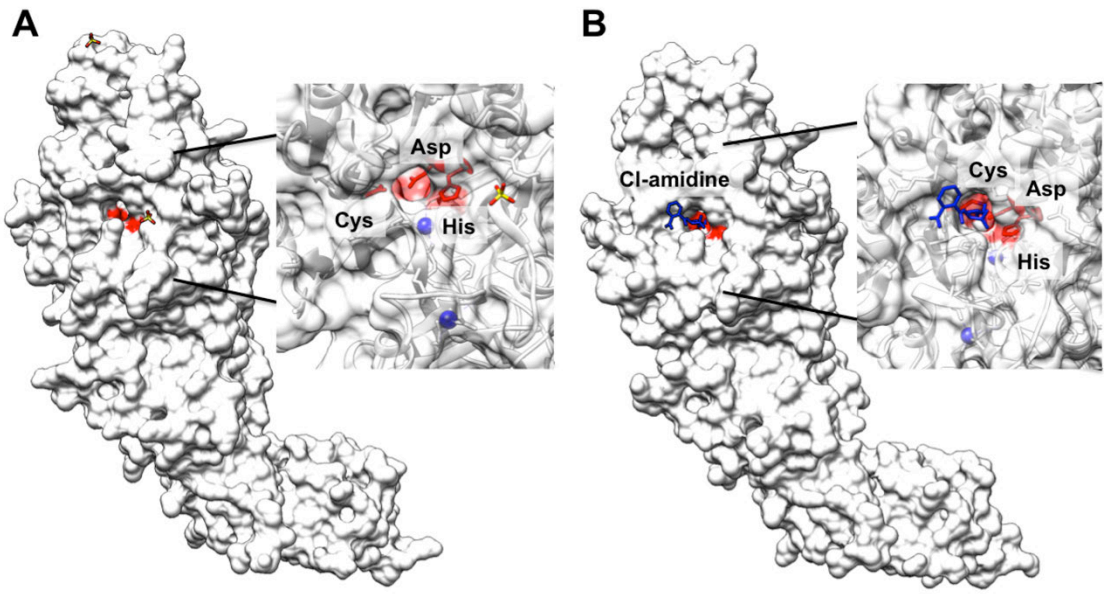


Figure 5.2: Crystal structures showing PADI4 selection designs for inhibitors and activators. **A:** Crystal structure of PADI4 soaked in 10mM $CaCl_2$ (pdb: 1wd9) to maintain it in the active conformation visualized in Chimera with protein surface displayed. Active site triad (Asp, His, Cys) residues are coloured in red showing the structured active site cleft. **B:** Crystal structure of PADI4 soaked in 10mM $CaCl_2$ to maintain it in the active conformation, with the active site blocked by the irreversible inhibitor Cl-amidine (pdb: 3b1t) visualized in Chimera with protein surface displayed. Active site triad (Asp, His, Cys) residues are coloured in red showing the structured active site cleft. The active site cysteine is covalently ligated to Cl-amidine (coloured in dark blue). Figures were prepared in Chimera using PDB structures 1wd9 and 3b1t.

The selections were performed by Dr Louise Walport in the lab of Prof Hiroaki Suga at the University of Tokyo. Candidate peptides were sequenced and synthesized by SPPS and sent in lyophilized form for me to test in downstream assays at the MRC Human Genetics Unit, Edinburgh.

5.3 Inhibitor peptides

5.3.1 Design for an *in vitro* screening assay for PADI4 inhibition

A citrullination assay was required to screen the candidate peptides *in vitro* as well as for other uses described elsewhere in this PhD thesis. Given that PADI activity has been shown to be robust to the presence of detergent in previously published assays[3,37], a citrullination assay using cell lysates was

developed. In *Darrah et al*., undifferentiated HL60s were lysed in 1% NP-40, sonicated and cleared by centrifugation before incubation with recombinant PADI enzymes, providing the enzymes with a wide range of substrates[37]. These were detected using western blotting with a protocol and antibody that detects chemically modified citrullinated products (Mod-Cit). This implied that PADI enzymatic activity might be sustained in the presence of as much as 1% NP-40 detergent. In *Lewis et al*.[3], a similar type of assay was used to test the selectivity of different PADI orthologues. Stable HEK293 cell pools expressing FLAG-PAD1, FLAG-PAD2, FLAG-PAD3 or FLAG-PAD4 were engineered and lysed in 0.4% NP-40 with the same detection method by Western blotting against anti-modified citrulline.

Based on these two approaches, I set up a lysate assay using the human PADI4 expressing mES cell line described in Chapter 4 (PADI4-stable mES cells)[38]. A control mES cell line that contains the empty vector was also included in the assays (Control-stable mES cells). Cells were washed in PBS with 0.5 mM EDTA and lysed by adding a lysis buffer containing 1% NP-40 directly to the plate and scraping the cells (Chapter 2.2.6.1). A relatively high concentration of NP-40 was selected to disrupt the nuclear membrane more efficiently. Benzonase was added and the lysate passed through a 25 G needle to shear and digest genomic DNA, disrupt the chromatin, and release histone substrates and chromatin-bound PADI4. The lysate was cleared by centrifugation, pooled and separated in equal volumes into different tubes preincubated on ice. Additional components such as calcium chloride or the cyclic peptide candidates may be added at this point to the lid of the tube made up to an identical total volume, and the assay initiated by a quick bench top centrifuge spin. Boiling the tubes at 95˚C quenches the assay. The activity of PADI4 can then be tested under a carefully controlled calcium concentration; other recombinant factors, inhibitors or cofactors can be added directly to the assay as required.

The lysate assay design provides several advantages. As the protein components are all produced by the cultured cells, there is no need to express, produce, and purify batches of recombinant enzyme or substrate. Since both enzyme and substrate are produced in the same starting cell lysate mixture, the relative concentrations of each will be exactly the same between each assay condition such that differences in total amounts of substrate/enzyme are kept to within a single pipetting error measurement and the ratio between the two will be identical. This assists in ensuring consistency in medium-throughput format when many samples are run for detection by Western blotting. Similarly, all the sample preparation for running by Western blotting occurs prior to the start of the assay such that no variability is introduced between conditions after the assay.

Figure 5.3 shows the results of the assay set-up. No activity is detected from the control cells in any condition showing that all activity under these experimental conditions is due to the exogenous human PADI4 enzyme. In PADI4-stable mES cells, a small amount of histone 3 substrate is endogenously citrullinated by human PADI4 in the resting cell state prior to assay incubation (Lane 12), but no additional citrullination can be detected after incubation in the absence of calcium during the course of the assay (Lane 7). Human PADI4 enzyme therefore shows the expected absolute requirement for calcium in this assay. In 5 mM Ca^{2+}, a large extent of H3Cit can be detected after the assay incubation showing a wide maximum detection range for inhibition. This level of citrullination is clearly reduced in conditions where Cl-amidine and GSK484 have been added. In neither Lane 10 (Cl-amidine) nor Lane 11 (GSK484) is H3Cit totally inhibited to background levels (Lane 7), showing that these inhibitors are not saturating in this assay even at their respective recommended concentrations. Full inhibition in this assay is a stringent test for candidate inhibitors (Figure 5.3). At their recommended concentrations, Cl-amidine (at 200 μM) can be seen to be more effective than GSK484 (5 μM). This indicates that the lysate assay is sensitive to differently effective inhibitors, which gives plenty of scope to

compare the efficacy of inhibition by any potential peptide candidates. A couple of additional observations were interesting: NPM1 appears to shift to higher apparent molecular weight in the activated conditions, most likely as a result of citrullination and autocitrullination respectively. As the mass change resulting from citrullination is small (0.98Da), the concomitant loss of charge is likely to result in the differing behavior under electrophoresis conditions in affecting protein motility. The reported citrullination site of NPM1 (R197) is not contained within the N-terminal epitope of the antibody to NPM1 (the epitope is an 14-22 amino acid region within a synthetic peptide corresponding to N-terminal amino acids 6-55 of human NPM1) and does not appear to affect antibody binding as the loading appears to be equal between conditions. A similar effect appears to be occurring to human PADI4.

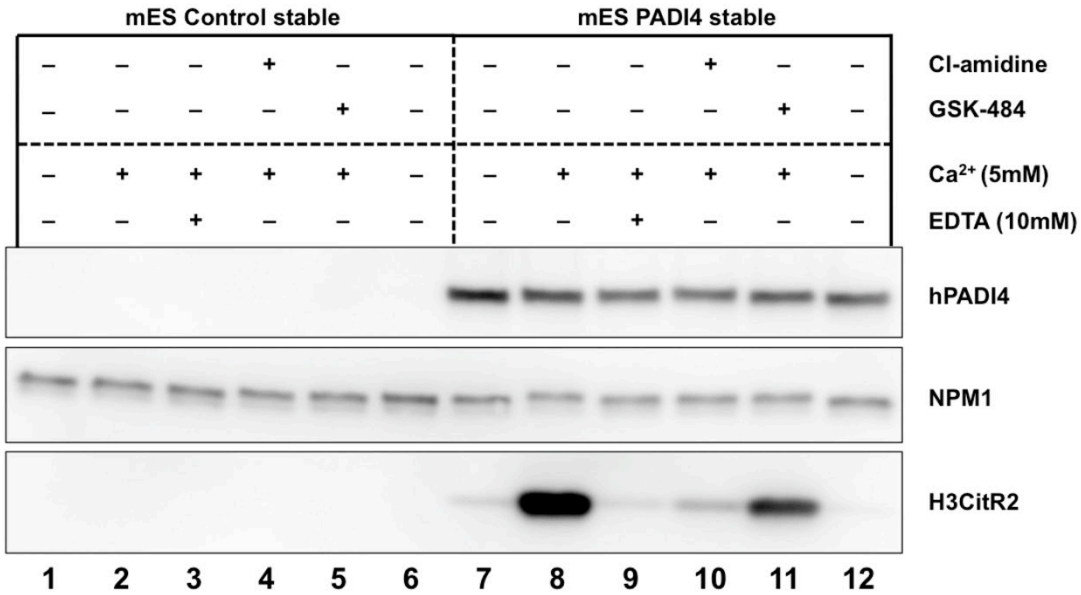


Figure 5.3: Initial set-up of the lysate citrullination assay to detect PADI4 activity. Immunoblot analysis of H3CitR2 and PADI4 of a lysate citrullination assay using Control-stable mES cells (Lanes 1-6) or PADI4-stable mES cells (Lanes 7-12). Lysates were incubated in the absence (Lanes 1 and 7) or presence of Ca^{2+} (5mM) (Lanes 2-5, 8-11) for 30 mins. Cl-amidine (200μM), GSK484 (100μM) or EDTA (10mM) were added at the start of the assay. Lanes 6 and 12 show cells lysed before the start of the assay incubation. Data are representative of n = 3.

5.3.2 Screening for PADI4 peptide inhibitor candidates

Candidate peptides (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 11b, and 14) at three different concentrations (200 μM, 20 μM, 2 μM) were then screened for PADI4 inhibition *in vitro*. Peptides were reconstituted from solid in pure DMSO at a high stock concentration (Chapter 2.1.18) such that DMSO concentrations were kept below 1% per assay condition. Control conditions always contained the same amount of DMSO as the peptide treated samples (labelled as “vehicle”). Spectrophotometry was used to measure the peptide concentrations using the predicted extinction coefficients and concentrations calculated using the Beer Lambert law (A = ε.c.l). This is likely to be more accurate than accurate mass measurements as different peptides in lyophilised salt form after SPPS will have unpredictably different counter ions that will affect their relative masses. Peptides or previously published inhibitors were added to the lid with calcium chloride, and assays begun by quick centrifugation. The peptide names in the screen were blinded both to the peptide sequence and to which method of selection (for inhibitors or activators) they had been enriched in. Although the majority of peptides (peptides 4, 5, 6, 7, 8, 9, 10, 11, 11b, and 14) did not show inhibition (Figure 5.4), two candidates (peptides 2 and 3) showed a clear inhibitory effect in the assay (Figure 5.4A) at 200 μM and 20 μM concentrations. Peptide 3 showed full inhibition to background levels even at 2 μM concentration (Figure 5.4A, final six lanes).

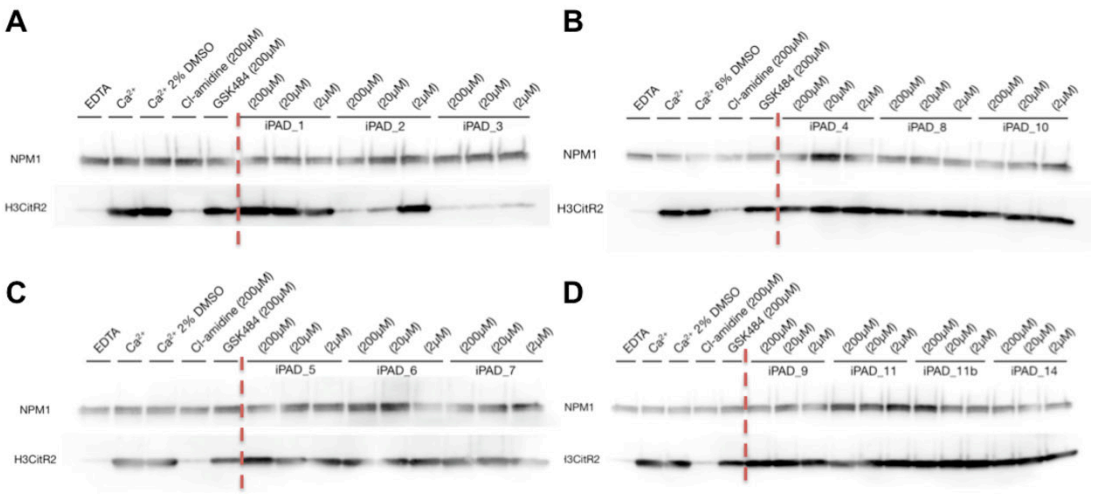


Figure 5.4: Screen of candidate cyclic peptide inhibitors. Immunoblot analysis of H3CitR2 of a lysate citrullination assay supplemented with vehicle, Cl-amidine (200μM), GSK484 (200μM) and serially diluted peptide # (200μM, 20μM, 2μM) in the presence of 5mM Ca^{2+} for 30 minutes. Data shown are from a single preliminary screening experiment; candidate peptides 2 and 3 from Panel **A** were taken forward for validation.

The assay was repeated at ten-fold serial dilutions of peptide 3 from 20 μM down to 2 nM (Figure 5.5). This experiment showed that Peptide 3 resulted in partial inhibition down to a concentration of at least 20 nM. Peptide 3 showed comparable inhibition to the best previously published reversible PADI4 inhibitor (GSK484) at a thousand-fold lower concentration (Figure 5.5).

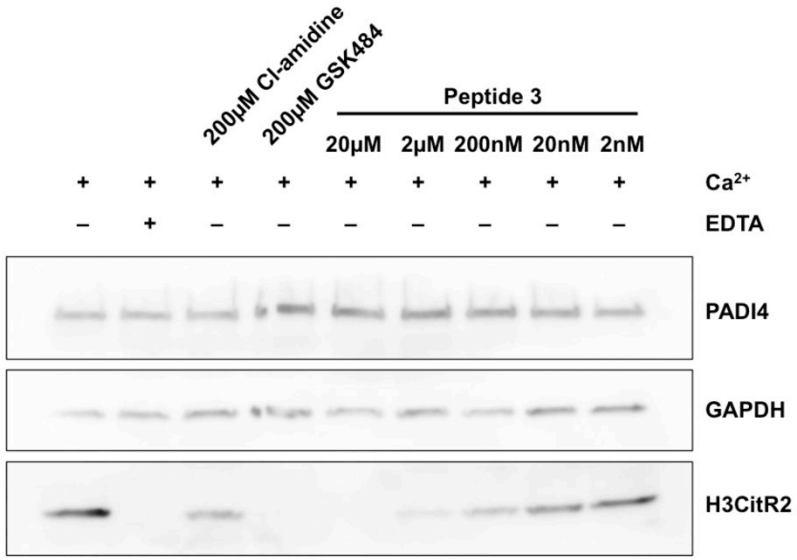


Figure 5.5: Peptide 3 potently inhibits PADI4 *in vitro*. Immunoblot analysis of H3CitR2 of a lysate citrullination assay supplemented with Cl-amidine (200μM), GSK484 (200μM) and serially diluted peptide 3 (20μM, 2μM, 200nM, 20nM, 2nM) in the presence of 5mM Ca^{2+} for 30 minutes. Data are representative of n = 2.

GSK484 has been previously published to have much a higher efficacy *in vitro* than observed in these assays[3], but is also reported to be five times less potent in a 2 mM calcium buffer than in the low calcium buffer for which its IC_{50}s were measured[12]. Given that GSK484 binds to a loop in the inactive calcium unbound form of PADI4, I hypothesized that a pre-incubation step with the enzyme might therefore increase the potency of GSK484 to more comparable levels to those published previously[3]. The assay was therefore

repeated such that the peptide or other inhibitors were added for 20 minutes prior to the addition of calcium chloride. This showed a marked increase in the efficacy of GSK484 at 100 μM (Figure 5.6). These data are consistent with the reported mode of action of GSK484 in targeting only the inactive form of PADI4. In addition, these data show that already activated PADI4 will continue to be active even in the presence of a very high concentration of GSK484 (Figure 5.6). The irreversible inhibitor Cl-amidine at 200 μM (Figure 5.6: Lane 3 versus Lane 10) as well as peptides 2 and 3 at 50 μM (Figure 5.6: Lanes 5 and 6 versus Lanes 12 and 13) were equally effective with or without pre-incubation steps, showing they can also inhibit the activated enzyme.

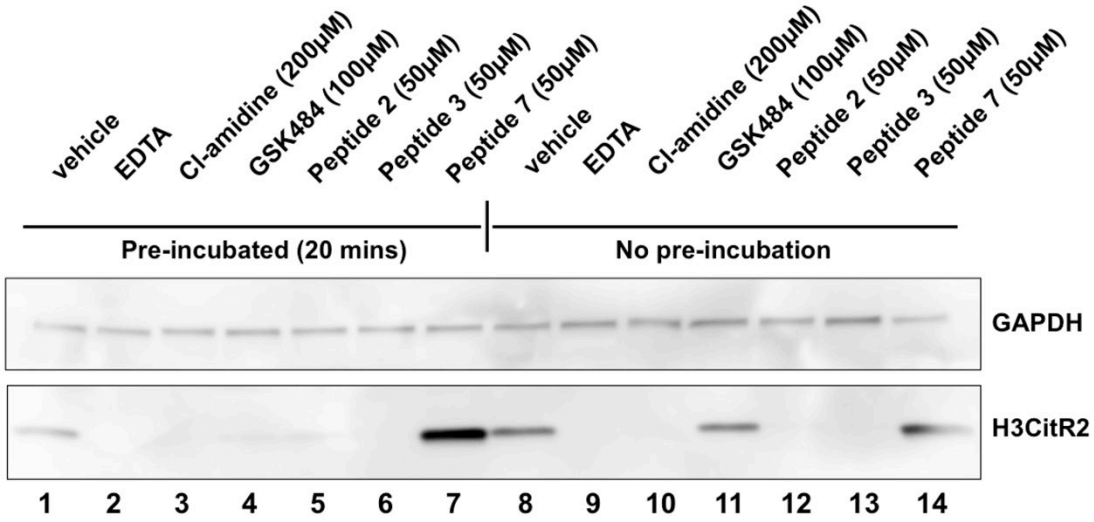


Figure 5.6: Comparison of peptides 2 and 3 for inhibition of PADI4 with GSK484. Immunoblot analysis of H3CitR2 of a lysate citrullination assay. In Lanes 1-7 lysates were preincubated on ice with vehicle, EDTA (10 mM), Cl-amidine, GSK484, peptide 2, peptide 3 or peptide 7 for 20 minutes, before $CaCl_2$ was added to begin the assay and the lysates incubated at 37˚C for 30 minutes. In Lanes 8-14, no preincubation step was performed and inhibitors or peptides were added concurrent to the addition of $CaCl_2$ and incubated for 30 minutes at 37˚C. Data are representative of n = 2.

5.3.3 Testing peptides on cells: inhibitors

The inhibitor peptides were then tested for their effect on cellular PADI4 (peptides 2 and 3) using conditions established in the previous chapter. Firstly, the activation condition of PADI4-stable mES cells treated for 6 hours

in KSR2i, was used to test the effect of the peptide inhibitors on cells (Figure 4.8A). As discussed in Chapter 4 (Section 4.2.5), this activation condition makes use of serum withdrawal and culture in knockout serum replacement (KSR), in combination with two specific small molecule inhibitors of the kinases GSK3β and MEK1/2 respectively (abbreviated as KSR2i) during a 6-hour treatment. KSR2i has precedent in the literature in increasing the efficiency of reprogramming somatic cells into induced pluripotent stem cells and in the establishment of ground state pluripotency[39,40]. Peptide 2 and 3 both reduced levels of H3Cit (Figure 5.7). H3Cit is taken as a proxy for PADI4 inhibition in this system. Cells were treated with this activation condition concurrently with four different peptide concentrations (75 μM, 15 μM, 3 μM, 0.6 μM) for 6 hours, with total DMSO concentrations kept constant across all treatments at 1% (Figure 4.8B and C). Peptide 3 showed efficacy at 75 μM with some efficacy at 15 μM and was observed to be more potent than peptide 2. It was concluded therefore that although both peptide 2 and 3 appear to be somewhat cell permeable, peptide 3 was more potent in all tested cases, both *in vitro* and in cells. To streamline efforts, peptide 3 was taken forward for validation and sequence optimization.

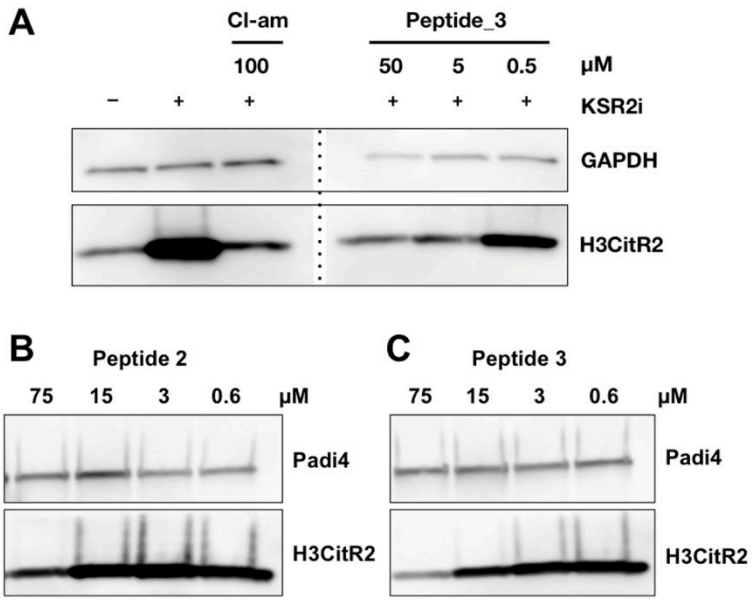


Figure 5.7: Peptide 2 and 3 inhibit PADI4 in cells activated by KSR2i. **A:** Immunoblot analysis of H3CitR2 of whole cell extracts of mES PADI4-stable cells after six hour treatment with serum and vehicle (Lane 1), KSR2i (Lane 2-6) in the presence of vehicle (lane 2), Cl-amidine 100μM (lane 3) or peptide 3 at 50μM, 5μM and 0.5μM concentrations (Lane 3-6). 2i

refers to the use of 1μM PD03250910 (inhibitor of MEK1/2) together with 3μM CHIR99021 (inhibitor of GSK3β)[58]. Dotted line indicates the removal of irrelevant lanes, from the same gel at the same exposure. Panels B and C show immunoblot of H3CitR2 of whole cell lysates extracted from PADI4-stable mES cells cultured in KSR2i for 6 hours in the presence of an increasing concentration of **B:** peptide 2 and **C:** peptide 3. In Panel A, data shown are representative of n = 3; in Panel B, data shown are from a single preliminary experiment; in Panel C, data shown are representative of n = 3.

Peptide 3 was then tested in a second activation condition identified in Chapter 4 (Section 4.2.6). In this second activation condition, activation of PADI4 was observed after use of the GSK3beta inhibitor (CHIR99021) after 45 minutes (without serum withdrawal or MEK1/2 inhibition) (Figure 4.9). Peptide 3 was pre-incubated with the cells for 2 hours and then CHIR99021 applied for 45 minutes (Figure 5.8). Full PADI4 inhibition occurred in the presence of 30 μM peptide 3 to the same level as control. Inhibition occurred dose dependently with some inhibition at 10 μM peptide, but no inhibition observed at 3 μM (Figure 5.8).

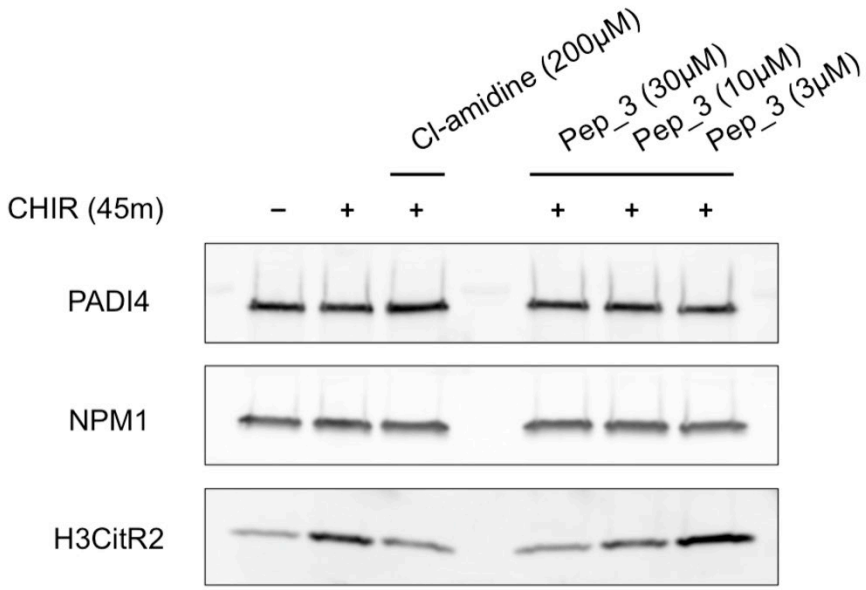


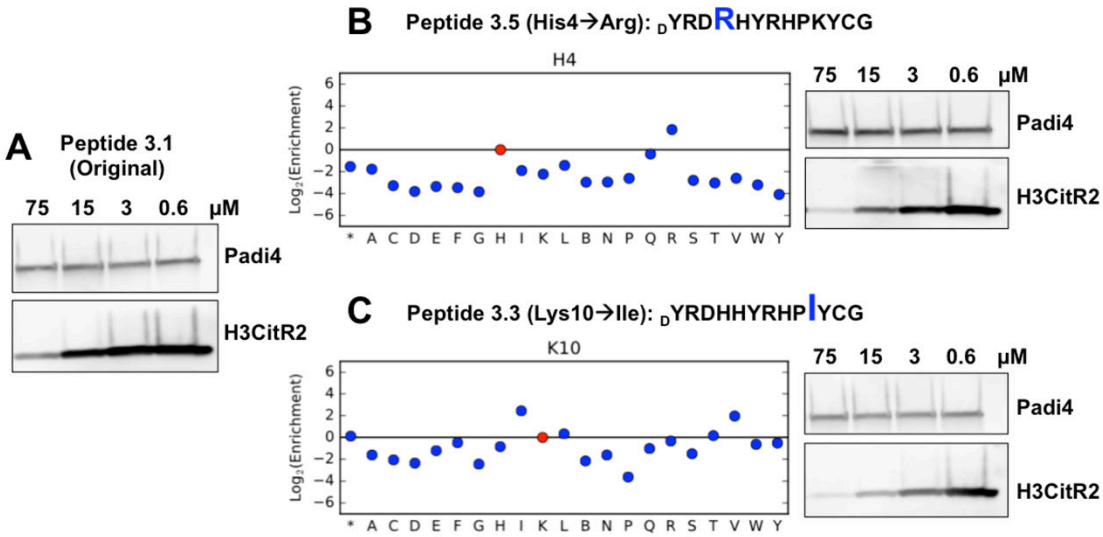
Figure 5.8: Peptide 3 inhibits PADI4 in cells activated by GSK3 inhibition. Immunoblot of H3CitR2 of whole cell lysates extracted from PADI4-stable mES cells pre-treated for 2 hours in the presence of an increasing concentration of peptide 3 (30μM, 10μM, or 3μM) before treatment with CHIR99021 at 3μM for 45 min. Data shown are representative of n = 3.

5.3.4 Optimising the peptide 3 sequence for cellular potency

We reasoned that it might be possible to optimize the peptide 3 sequence further for cellular potency. To do this, Dr Louise Walport performed a mutational scanning experiment with the RaPID selection methodology, using a strategy recently developed in the Suga lab[41]. Briefly, a new DNA library was prepared, using the parent peptide 3 sequence as a template, in which every residue was individually mutated to every other amino acid. Multiple members of each alternative sequence were included. The original library was then sequenced to identify what the original make-up of the library is at the start and then a single selection was performed with this library against the protein target, with long washes (3 x 12 hour washes) to ensure equilibrium binding is reached with the target protein[41]. The input and output libraries were then sequenced and compared to produce $\log_2$ ratios of enrichment for every mutated residue in the parent sequence. This in theory should be proportional to the K_D of the new sequences (provided equilibrium binding is achieved during selection). In addition, a codon for backbone N-methylated alanine was used, so the $\log_2$ ratio was obtained for N-methyl alanine over alanine for each residue in the parent sequence is included in addition. From the comparison of enrichment of N-methyl alanine against alanine, Dr Walport hypothesized that it may be possible to deduce which residues may be well-tolerated for backbone methylation.

Dr Walport then synthesized a first round of peptide variants for me to test in citrullination assays on cells. After this first set of results, a second round of optimization was undertaken, including a myristoylated variant of peptide 3 and both Arg-Cit mutants of the parent sequence. The latter variants were selected based on the results of the Arg->N and Arg-W mutants from the first round (which at these positions revealed enrichment) and from the results of the SPR data (Figure 5.15) where peptide 3 mutation from arginine to citrulline still maintained nanomolar binding affinity. A handful of double mutants were also prepared. Citrullination assays were then performed on the second round of optimized sequences.

All assays were performed blinded to the sequence name or identity (Peptides 3.2, 3.3, 3.4, 3.5, 3.6, 3.7, 3.8, 3.9). Figure 5.9A-B shows the efficacy of two of the synthesized peptide 3 variants used in cell assays (right panels, Figure 5.9A-B) alongside the panel of log_2 ratio of enrichment obtained from mutational scanning for that residue (left panels, Figure 5.9A-B). As before, Dr Walport then obtained SPR data for the new variants. The efficacy of all tested peptide 3 variants is summarized in a table (Figure 5.9C) together with the SPR data obtained by Dr Walport.



Name	Substitution	Sequence	Improvement over Peptide 3	K_D by SPR
PAD4_3	–	dYRDHHYRHPKYCG	n/a	4.15E-09
PAD4_3.2	Myristoylation	dYRDHHYRHPKYCGSβAK(myr)	–	7.73E-09
PAD4_3.3	K10I	dYRDHHYRHPIYCG	+	3.23E-09
PAD4_3.4	H4R, K10I	dYRDRHYRHPIYCG	+	4.00E-09
PAD4_3.5	H4R	dYRDRHYRHPKYCG	+	7.50E-09
PAD4_3.6	H8H-N_Methyl, K10I	dYRDHHYRH_{Me}PIYC	+	3.25E-07
PAD4_3.7	H8H-N_Methyl	dYRDHHYRH_{Me}PKYC	–	2.35E-07
PAD4_3.8	R2Cit	dYcitDHHYRHPKYCG	(+)	2.00E-09
PAD4_3.9	R7Cit	dYRDHHYcitHPKYCG	+	5.91E-09

Figure 5.9: Optimising inhibitor peptide sequences. A RaPID selection against human PADI4 was performed with a DNA library containing a known number of copies of variants of the initial peptide 3 sequence (YRDHHYRHPKYCG) where every amino acid was singly permuted to all other possible amino acids along with a known number of copies of the

original sequence. Several rounds of selection were performed with long washes after which all barcodes were sequenced. The log_2(ratio) for variant sequence enrichment over the original sequence was plotted (y-axis) for each mutated residue (x-axis) in a separate graph for every residue in the original peptide 3 (one graph of log_2(ratio) enrichments per residue in the original sequence). A plot for all possible single substitutions are shown in the middle panel for **B:** the 4th residue (His) of peptide 3 and **C:** the 10th residue (Lys) of peptide 3. The right hand panels show immunoblot analysis of whole cell lysates extracted from PADI4-stable cells cultured in KSR2i for 6h, in the presence of an increasing concentration of **A:** the original peptide 3 sequence, as well as two synthesized peptide variants **B:** peptide 3.5 (variant His4 to Arg) and **C:** peptide 3.3 (variant Lys10 to Ile). Data are representative of n = 2. **D:** Table showing sequences of the peptides tested including all variants of the original peptide 3 tested with results of tests in cells (column 4) and SPR binding data obtained from experiments by Dr Walport (column 5). Dr Walport performed RaPID screen and synthesized all cyclic peptides.

5.4 Activators

5.4.1 Screening for PADI4 peptide activator candidates

The citrullination lysate assay to screen for candidate inhibitors was then adapted to screen for putative PADI4 peptide activators. To do this, the citrullination lysate assay was performed at a single concentration of peptide (100 μM) at three different calcium concentrations (100 μM, 500 μM, 1 mM) for 45 minutes. This Ca^{2+} range spans the absolute requirement for calcium of human PADI4 on the histone 3 substrate (Figure 4.2A)[7,42]. I hypothesized that a candidate activator might act to lower the calcium dependence of PADI4 (in other words increase the calcium sensitivity) in this assay as compared to a vehicle treated condition.

The assay was then performed to screen all of the peptide candidates for their ability to activate PADI4 *in vitro* (peptides 1, 4, 5, 6, 7, 8, 9, 10, 11, 11b and 14). The peptide names were blinded both to the peptide sequence and to the method of selection (for inhibitors or activators) from which they were discovered. Peptide 2 and 3 were not included, as they had already showed efficacy as inhibitors. Although the majority of peptides (peptides 1, 4, 5, 6, 7, 8, 9, 10 and 11b) did not show any clear activating effect at the lower calcium concentrations (100 μM and 500 μM), excitingly, two candidates (peptides 11

and 14) increased the amount of H3Cit detected at 100 µM and 500 µM $CaCl_2$ over vehicle treatment at the same calcium concentration and resulted in a comparable extent of citrullination to that observed at 1 mM $CaCl_2$ under vehicle treatment (Figure 5.10).

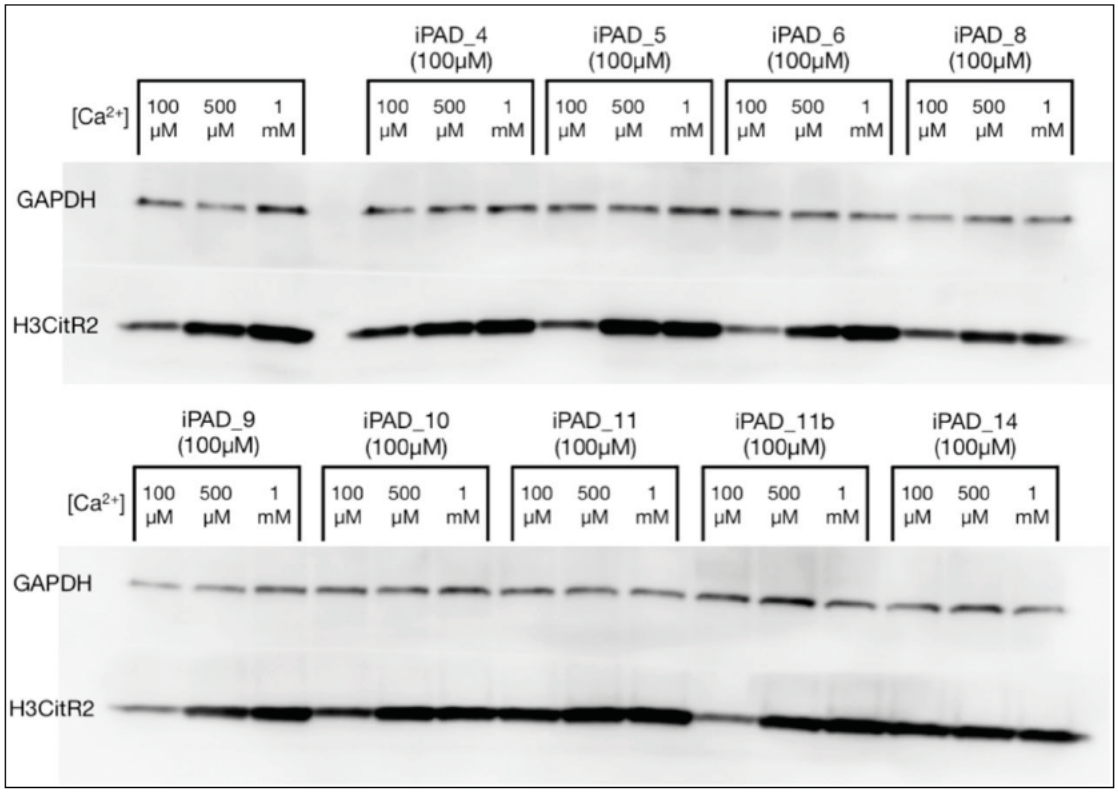


Figure 5.10: Screen of candidate cyclic peptide activators. Immunoblot analysis of H3CitR2 of a lysate citrullination assay. **A:** Vehicle or 100µM peptide # (4, 5, 6, 8, 9, 10, 11, 11b, 14) were added with serially diluted $CaCl_2$ (100µM, 500µM or 1mM) and incubated with the lysate for 45 minutes. Data shown are from a single preliminary screening experiment; candidate peptides 11 and 14 were taken forward for further validation.

5.4.2 Characterising candidate peptide activators

These two promising activating peptides (peptide 11 and peptide 14) were taken forward for further testing at a wide range of serial dilutions of calcium chloride and for a shorter incubation of 30 minutes (0, 15.5, 31, 62.5, 125, 250, 500 and 1000 µM). Another peptide (peptide 9) was also included as a negative control peptide as it had shown no effect in either of the first inhibition or activation *in vitro* screens (Figure 5.11B). A 100 µM concentration of peptide was used in each condition. A further control using

vehicle was also performed (Figure 5.11A). Using control Peptide 9 or vehicle, no activity could be detected at 500 µM Ca^{2+} or lower during the 30-minute incubation, but robust activity was detected in the presence of a minimum of 1 mM Ca^{2+} (Figure 5.11 A and B, left hand panels). By contrast, the addition of either peptide 11 or peptide 14 results in activity of PADI4 at the lower 500 µM concentration of calcium ions (Figure 5.11 A and B, right hand panels). This shows that the presence of peptide 11 or 14 lowers the calcium dependence of human PADI4 and thereby drives enzyme activation *in vitro*. The control peptide ensures that an unknown artefactual cofactor introduced in the synthesis of the peptides (such as in the SPPS steps) was not responsible for the observed activation effect. This indicates that the specific sequences of Peptides 11 and 14 are responsible for the effect, although it is possible that the activation effect could be mediated indirectly through binding to a different component in the lysate.

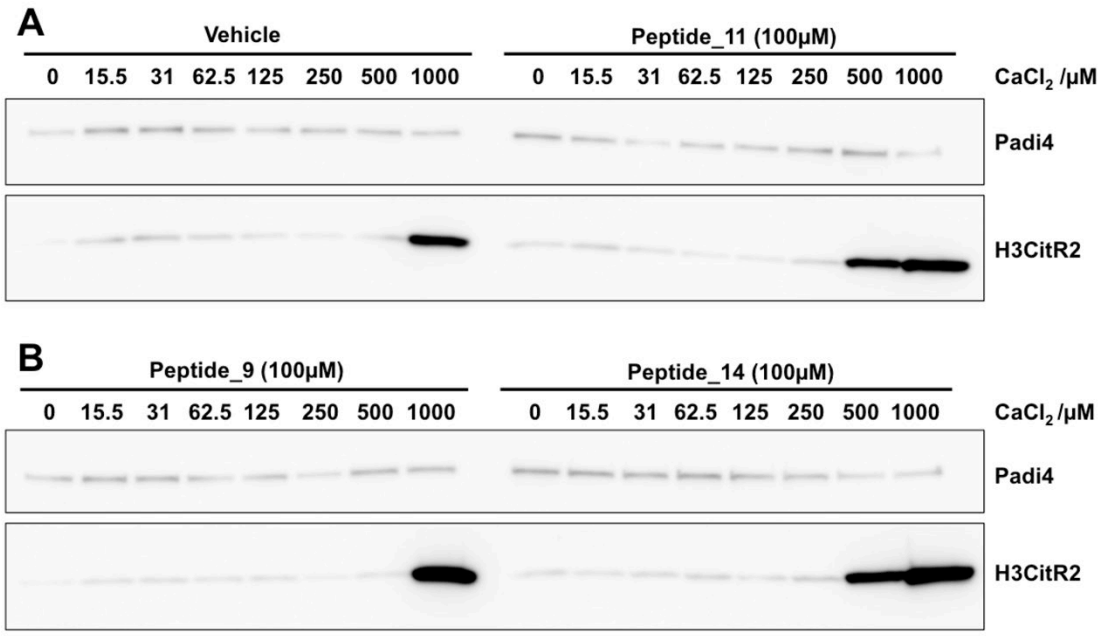


Figure 5.11: Peptides 11 and 14 activate PADI4 *in vitro*. Immunoblot analysis of H3CitR2 of a lysate citrullination assay. **A:** Vehicle (Lanes 1-8, left hand side of panel) or 100µM peptide 11 (Lanes 9-16, right hand side of panel) were added with serially diluted CaCl2 and incubated with the lysate for 30 minutes. **B:** Control peptide 9 (Lanes 1-8, left hand side of panel) or 100µM peptide 14 (Lanes 9-16, right hand side of panel) were added with serially diluted $CaCl_2$ and incubated with the lysate for 30 minutes. Data are representative of n = 2.

A further preliminary assay was also performed at a longer incubation of 90 minutes. During the 90 minute incubation, activity could be detected in as low as 250 μM Calcium (Figure 5.12B). Additionally, in each case, peptide 14 elicited a greater activating effect than peptide 11 (Figure 5.12B). At this long incubation, activity in vitro appeared to be saturated at 500μM $CaCl_2$. Further experiments will be need to explore further the limiting concentration of activating peptides and optimal in vitro incubation time.

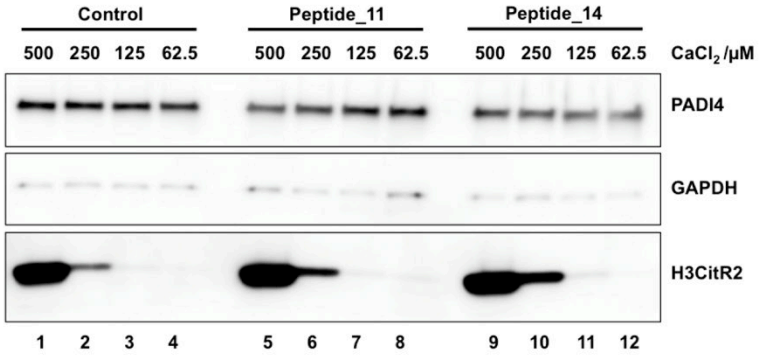


Figure 5.12: Characterising peptides 11 and 14 *in vitro*. Immunoblot analysis of H3CitR2 of a lysate citrullination assay. Vehicle (Lanes 1-4), 50μM peptide 11 (Lanes 5-8), or 50μM peptide 14 (Lanes 9-12) were added with serially diluted $CaCl_2$ (500μM, 250μM, 125μM, 62.5μM) and incubated with the lysate for 90 minutes. Data shown for these additional conditions are from a single preliminary experiment.

5.4.3 *In vitro* experiments to analyse the activating peptides on recombinant proteins

These experiments so far have shown the capacity for these peptides to activate endogenous protein from a lysate. We firstly wanted to test whether the peptides bound at the enzyme active site. Dr Walport therefore performed an assay using a biotin conjugated F-amidine molecule developed in the Thompson group[43] (Figure 5.13A). F-amidine, unlike Cl-amidine, reacts only with the activated calcium-bound conformation of PADI4[43]. Peptides 1, 2, 3, 7, 11, and 14 were first incubated with PADI4 in the presence of calcium, subsequently incubated with the biotin conjugated F-amidine compound, and then run on an SDS PAGE gel with detection by Streptavidin-HRP. Preincubation with peptides 2 and 3, but not 1, 7, 11, or 14, abrogated Streptavidin-HRP signal. This confirms that the inhibitor

peptides 2 and _3 clearly bind at the active site of PADI4 as they prevent biotin-conjugated F-amidine reacting with the active site of active PADI4 (Figure 5.13B). The experiment lastly suggests that peptide 11 and 14 activate by binding allosterically as suggested by their discovery from the second selection method, as their binding does not block access to the active site.

We also considered that the calcium sensitivity of PADI4 may be increased further from the tests performed using lysates if recombinantly purified proteins only were used. Dr Walport therefore repeated the assay for peptides 11 and _14 at different calcium concentrations (0, 50 μM, 75 μM, 100 μM, 125 μM, 150 μM, 250 μM Ca^{2+}). We hypothesized that the activating peptides might enable an active conformation at lower concentrations of calcium. Streptavidin-HRP signal was observed down to 50 μM Calcium in the presence of peptides 11 and 14, where no signal could be detected in the absence of peptide. The intensity of Streptavidin-HRP signal was increased in the peptide 11 and 14 conditions over calcium only up to 150 μM, and at 250 μM the Streptavidin-HRP was saturating with equal signal in presence and absence of peptide. This shows that peptides 11 and 14 allow stabilization of the active PADI4 conformation down to 50 μM Ca^{2+} when using recombinantly purified protein. Discrepancies in calcium dependencies between real substrates such as histone 3 and BAEE (on which F-amidine is based) have been published and may explain the higher absolute calcium requirement on calcium which was observed in the *in vitro* cell lysate assays. This also helps rationalize how longer incubations of peptide with enzyme and substrate *in vitro* may nonetheless enable activity on histone substrate at a lower calcium dependency of PADI4 (Figure 5.13C). This is strong additional support that Peptides 11 and 14 activate PADI4 by allosteric binding; their presence increases reaction of the probe at the active site. Presence of peptides 11 and 14 thereby mediate formation of the active conformation of PADI4 with the active site available for chemical reaction with the probe at a reduced calcium ion concentration.

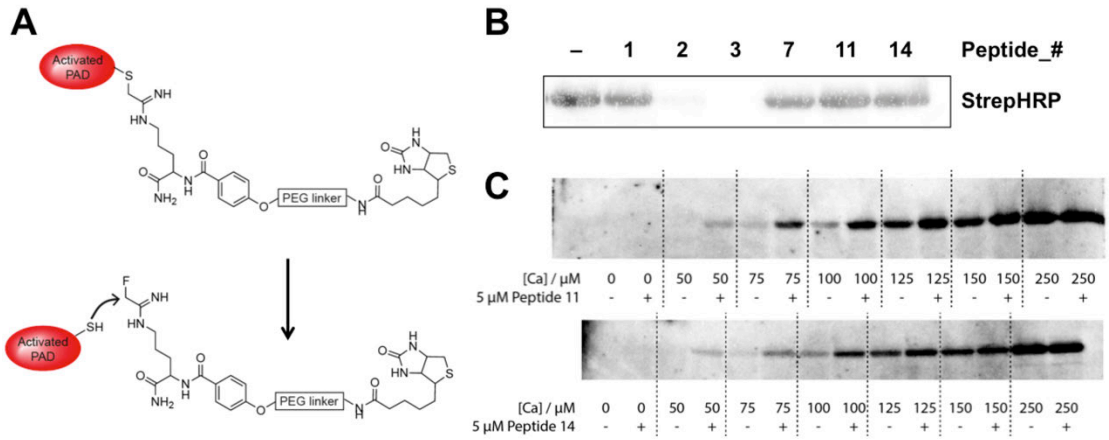


Figure 5.13: Testing peptides for active site binding and generating the active PADI4 conformation. **A:** A biotinylated_F-amidine reagent developed by the Thompson group[1,43]. The probe reagent (Bio-F-amidine) reacts irreversibly with the active site cysteine (Cys645) only if human PADI4 is in the active conformation, but is not reactive with the inactive conformation[43]. This reagent can be subsequently visualized by western blot using streptavidin conjugated to horseradish peroxidase (Streptavidin-HRP) and used to detect the presence of active PADI4. It was synthesized and used here with a PEG linker between biotin and F-amidine. **B:** Peptide 2 and 3, but not peptides 1, 7, 11, or 14, directly bind at the active site. Vehicle or peptide 1, 2, 3, 7, 11, 14 were preincubated with human PADI4 in the presence of Ca2+ before incubating with Bio-F-amidine. Detection was then performed with Streptavidin-HRP by western blotting at the molecular weight of PADI4. The lack of signal in lanes preincubated with either peptide 2 or peptide 3 indicate active site protection from Bio-F-amidine. **C:** Peptide 11 and 14 lower the calcium requirement to form active conformation of PADI4. Vehicle (–) or peptide 11 (+) (top panel), or vehicle (–) or peptide 14 (+) (bottom panel), were incubated with PADI4 at increasing concentrations of calcium before incubating with Bio-F-amidine and detection by western blot with Streptavidin-HRP. Experiments in this figure were performed by Dr Walport.

5.5 Testing activating peptides on cells

Given the promise shown by the inhibitor peptides on cells, I then decided to test the peptide activators directly on mouse embryonic stem cells expressing human PADI4 cultured in serum (Figure 5.14). Treatment with peptide 11 or peptide 14 activated PADI4 in these mouse ES cells without any additional activating stimulus. In serum, clear activation could be observed at 75 µM peptide 11 and as low as 30 µM peptide 14 (Figure 5.14).

Preliminarily, I then employed the peptides in combination with the KSR2i activation condition developed in Chapter 4 (Section 4.2.5, Figure 4.8). In combination with KSR2i the peptides appeared to activate at a concentration as low as 15 μM, suggesting that the modes of activation may work in tandem (Figure 5.14C) and clearly merits further repeats. It will be particularly exciting to test these activating peptide reagents in biological contexts and this is returned to in the discussion.

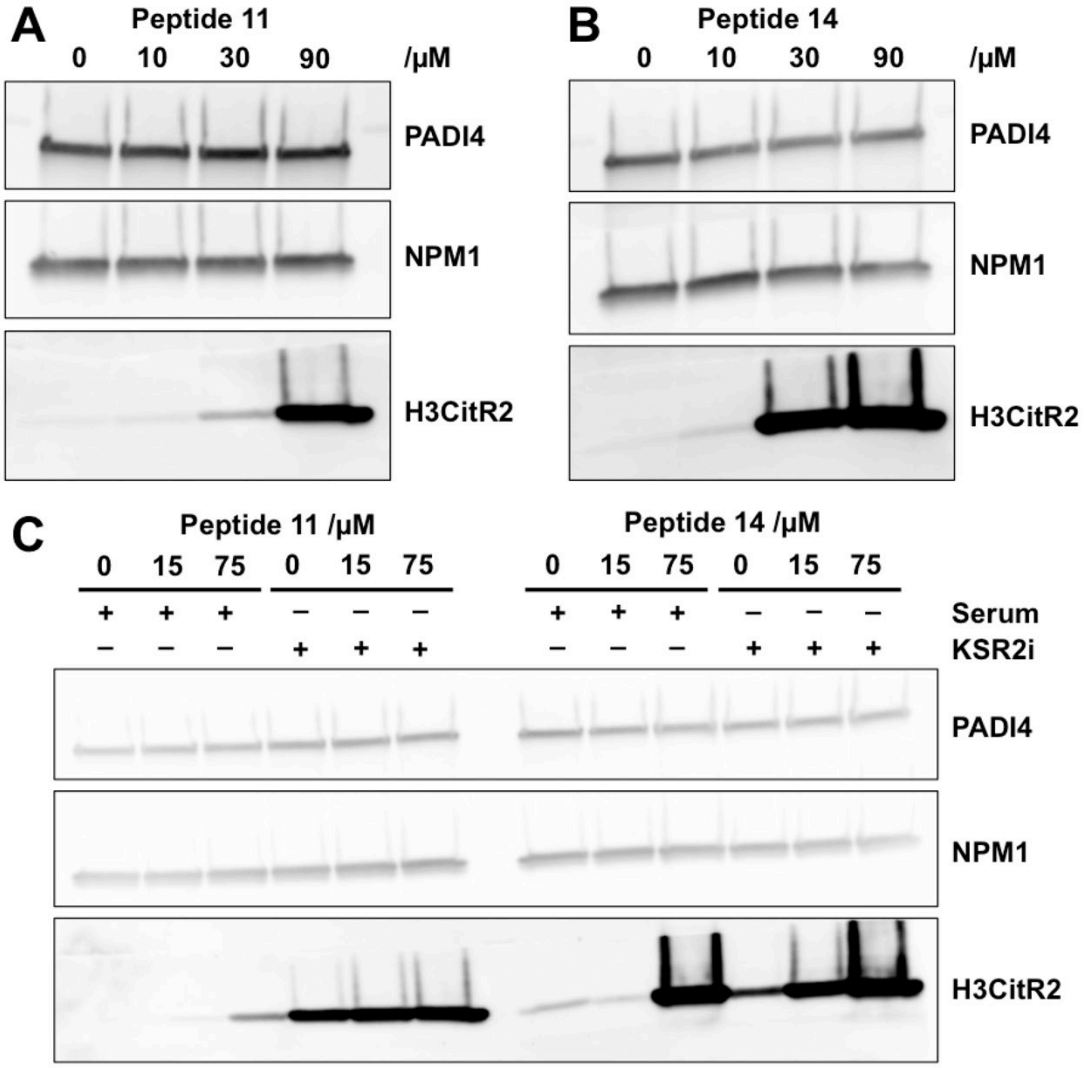


Figure 5.14: Peptides 11 and 14 activate PADI4 in cells. **A and B:** Immunoblot of H3CitR2 and PADI4 of whole cell lysates extracted from PADI4-stable mES cells treated with an increasing concentration (0 μM, 10μM, 30μM, 90 μM) of **A:** peptide 11 or **B:** peptide 14. **C:** Immunoblot of H3CitR2 and PADI4 of whole cell lysates extracted from PADI4-stable mES cells treated in serum (Lanes 1-3 and Lanes 7-9) or in KSR2i (Lanes 4-6 and 10-12) for 6 hours and incubated with an increasing concentration (0 μM, 15 μM, 75 μM) of peptide 11

(Lanes 1-3 and 4-6) or peptide 14 (Lanes 7-9 and 10-12). Data are representative of n = 3, except for treatments in KSR2i which are shown from a single preliminary experiment.

5.6 Peptide binding affinity measurements by SPR and binding mode

Based on this promising data *in vitro* and in cells, direct binding affinity measurements were then acquired by SPR at the University of Tokyo by Dr Walport under three different conditions: 1) without calcium, 2) with calcium and 3) with calcium and Cl-amidine preincubation. Given peptides 1, 2 and 3 all contain arginine residues, binding measurements were also taken for peptides where this arginine had been substituted for citrulline. We hypothesized that these peptides may inhibit by acting as substrates. Key results are summarized in Figure 5.15. Consistent with the *in vitro* inhibition and activation screens, peptides 2 and 3 bind very potently in calcium containing buffer (~35 nM and below 10 nM) but do not bind significantly in a no calcium-containing buffer or in a calcium buffer where PADI4 had been preincubuated with Cl-amidine. This is good additional evidence that they bind at the active site of PADI4 in the calcium-bound conformation of the enzyme. Peptides 2 and 3 do not bind the inactive, unstructured cleft and cannot bind if Cl-amidine is already occupying the active site. Peptide 3 is more potent at binding than peptide 2, which is consistent with its increased potency in the *in vitro* inhibition assay. Peptides 11 and 14 bind tightly in calcium buffer and to approximately the same extent whether or not the active site was blocked with Cl-amidine. Peptides 11 and 14, by contrast with peptides 2 and 3, do not bind PADI4 in the absence of calcium. Taken together, this is strong evidence that they interact allosterically and indicates that they may activate by binding to and stabilizing the calcium-bound enzyme conformation.

	No calcium buffer	Calcium buffer	Calcium buffer + ClAmidine preincubation	Arg to Cit substitution (Ca buffer)
PAD4_1	~160 nM	~130 nM	~200 nM	~1000 nM
PAD4_2	> 1000 nM	~35 nM	> 1 uM	~350 nM
PAD4_3	> 1000 nM	<10 nM	> 1 uM	~50 nM
PAD4_7	~ 10 nM	~ 25 nM	~ 25 nM	N/A
PAD4_11	> 1000 nM	~500 nM	~ 300 nM	N/A
PAD4_14	> 1000 nM	~ 800 nM	~ 500 nM	N/A

Figure 5.15: Peptide binding affinity measurements by SPR. Table showing K_D values obtained from SPR experiments performed by Dr Walport. Peptides 1, 2, 3, 7, 11, 14 were

tested for binding to human PADI4 in a no calcium buffer (column 1), high calcium buffer (column 2), or in a high calcium buffer in the presence of Cl-amidine (column 3). Peptides 1, 2 and 3 all contain an arginine residue within the cyclic lariat sequence, which was substituted for citrulline. SPR data were also obtained for these substituted peptides (column 4). Experiments in this figure were performed by Dr Louise Walport.

5.7 Developing an affinity molecule: Peptide 7

Peptide 7 showed comparable high affinity binding to PADI4 in buffers containing no calcium, in buffers containing calcium and to PADI4 that had been preincubated with Cl-amidine (Figure 5.14). In addition, after incubating peptide 7 with PADI4, biotinylated-F-amidine was still able to react at the active site showing peptide 7 does not bind the active site. Taken together, these data clearly demonstrate that peptide 7 binds away from the active site, at a location of the enzyme that is unaffected by calcium binding and therefore that is likely to be structured at all concentrations of Ca^{2+} (Figure 5.1, regions in white). Dr Walport and I reasoned that it therefore might make a very good candidate to explore as a reagent for affinity pulldown experiments. A major advantage of having the biotinylated peptide 7, in addition to the commercial antibody, is that it is likely to bind to PADI4 and pull it down in a different conformation from the antibody binding. It therefore may enable a more complete set of interacting partners to be identified and was a particularly useful reagent for the mass spectrometry experiments undertaken in Chapter 6.

I then elected to test whether this peptide might be suitable for use as an affinity molecule with Dr Christophorou. A biotinylated peptide 7 was synthesized by SPPS by Dr Louise Walport at the University of Tokyo. The identified peptide 7 sequence $_{D}$YYPKGSWGYKLFCG (underlined residues are those included in the cyclic lariat sequence) was synthesized with the inclusion of an additional serine, β-alanine, and biotinylated lysine. Two further variants were synthesized with even longer linkers: the first with an 11-carbon length chain (“Undec”) to replace the β-alanine residue, and a second with five β-alanine residues instead of one. The three peptides were

synthesized as lyophilized powders by Dr Walport. Based on lysis conditions I had established from antibody pull down (Chapter 6), Dr Christophorou and I then tested all three molecules with lysates from PADI4-stable mES cells according to two different protocols: 1) incubate the peptide overnight with the cell lysate, followed by a 2 hour incubation with streptavidin magnetic beads (as the pulldown step) or 2) pre-incubate the peptide with the streptavidin beads for 2 hours first, followed by overnight incubation with the cell lysate (Figure 5.16A). The second protocol produced a much higher efficiency pulldown for all three peptides (as had also been observed for my antibody pulldowns, Section 6.2). The original biotinylated peptide 7 was at least as effective as (and in fact somewhat more effective than) either of the longer linker peptides as detected by western blotting for PADI4 (Figure 5.16A, Lane 2 shows stronger signal than Lanes 3 or 4, and Lane 6 over Lanes 7 or 9). The second protocol additionally reduced the extent of background binding of PADI4 to the beads. The second protocol and original biotinylated peptide 7 were therefore taken forward (Lane 6) and showed good enrichment over input.

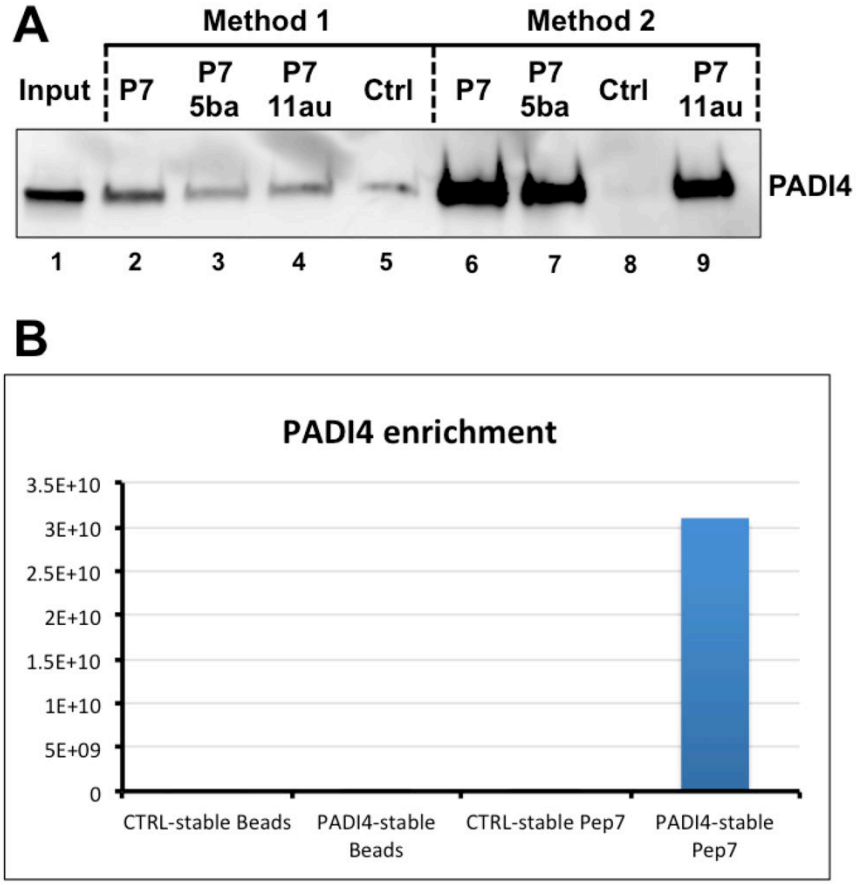


Figure 5.16: Biotinylated Peptide 7 isolates PADI4 from cells and is suitable for MS/MS analysis. **A:** PADI4-stable mES cells were lysed and incubated with biotinylated peptide 7, with no linker (Lanes 2 and 6), with a 5x beta-alanine linker (Lanes 3 and 7), with an 11au linker (Lanes 4 and 9) or with vehicle (Lanes 5 and 8). In Method 1, peptide and lysates were incubated overnight at 4˚C and then pulled down with streptavidin beads (Lanes 2-5). In Method 2, peptides were incubated with streptavidin beads first for 2 hours and then incubated with lysates overnight at 4˚C (Lanes 6-9). Detection was by western blot using an antibody to human PADI4. Data are representative of n = 2. **B:** Affinity pulldown followed by mass spectrometry analysis showing LFQ intensity of human PADI4 from analysis of proteomic data in MaxQuant. Pulldown was performed with beads alone or biotinylated peptide 7 from Control-stable mES cells or PADI4-stable mES cells. This experiment was performed with Dr Christophorou.

A preliminary experiment was then performed with Dr Christophorou to pull down PADI4 using biotinylated peptide 7 for analysis by quantitative mass spectrometry. Control cells and PADI4 stably expressing embryonic stem cells were grown and used for an affinity pulldown with biotinylated peptide 7

and streptavidin coated magnetic beads. Four conditions were assessed: 1) control cells pulled down with beads alone, 2) PADI4 cells pulled down with beads alone, 3) control cells pulled down with peptide 7, and 4) PADI4 cells pulled down with peptide 7. Beads were washed in 50 mM Ammonium bicarbonate and submitted for analysis by the IGMM Mass spectrometry facility. Briefly, dry beads were digested with trypsin and peptides purified by Liquid chromatography using C18 STop And Go Extraction (STAGE) tips. MSMS analysis was performed on a Thermo Q-Exactive Plus machine and data analysis was performed using MaxQuant. LFQ intensity for PADI4 protein across the four conditions is presented in Figure 16B. 27 different PADI4 tryptic peptides were identified and PADI4 was the second highest intensity protein identified in the peptide 7 pulldown from PADI4-stable cells (second to Actin). Actin was 20-fold enriched in peptide 7 pulldowns over beads in both PADI4-stable and control conditions implying it may additionally be pulled down by the peptide and not pulled down together with PADI4. Enrichment for PADI4 peptides was found only in the PADI4-stable condition pulled down using peptide 7: an 8000 fold increase in LFQ intensity was observed over beads alone and a 20 000 fold increase in LFQ intensity was observed over a peptide 7 pulldown from mES Control-stable cells. ~50 proteins were additionally highly enriched (>50 fold enrichment in LFQ intensity) in the peptide7 pulldown from PADI4-stable cells, including an identified interactor of PADI4 from BioGrid (ANXA4) suggesting other proteins could be co-pulled down.

5.8 Development of control peptides

It would be very desirable to have negative control peptides that are inactive for future use of these reagents in biological contexts as a control for the addition of cyclic peptides to cells (such as peptide 9, Figure 5.11A). With Dr Walport, a series of scrambled control peptides were therefore designed for peptide 3, 7, 11 and 14; these have the same amino acid content but with a disrupted sequence that in theory ought to render them inactive with respect to binding PADI4.

I started by testing the scramble inhibitor peptide 3, while I was testing the optimized peptide 3 sequences. Very unfortunately in preliminary experiments to test this on cells, the scrambled peptide also appeared to inhibit PADI4 in cells (Figure 5.17B). I therefore tested it in the *in vitro* citrullination lysate assay where it did not inhibit. These data on their own are hard to reconcile, and need to be repeated carefully. Given the data derive from a single preliminary experiment, technical problems with the experiment cannot be ruled out. If the effects, however, are validated then a possible, if very unfortunate, hypothesis is that the scramble peptide 3 is similarly cell permeable (it has the same amino acid composition) and affects the activation pathway in an unpredictable way, but without inhibiting PADI4 directly. It will be useful to test the inhibitors for efficacy in a different mode of activation such as that caused by the calcium ionophore A23187.

Scramble peptides may not be the best negative control for the active peptides. As an alternative, several variant peptides were already tested that showed little effects on inhibition and therefore act already as good negative controls. For example, the 3.7 variant of peptide 3 is backbone N-methylated with no change to the sequence. This variant does not inhibit PADI4 in cells (shown in Figure 5.17C) suggesting it can be a good negative control for peptide inhibition. Dr Walport had also confirmed that this variant bound PADI4 100-fold less strongly by SPR (Figure 5.9).

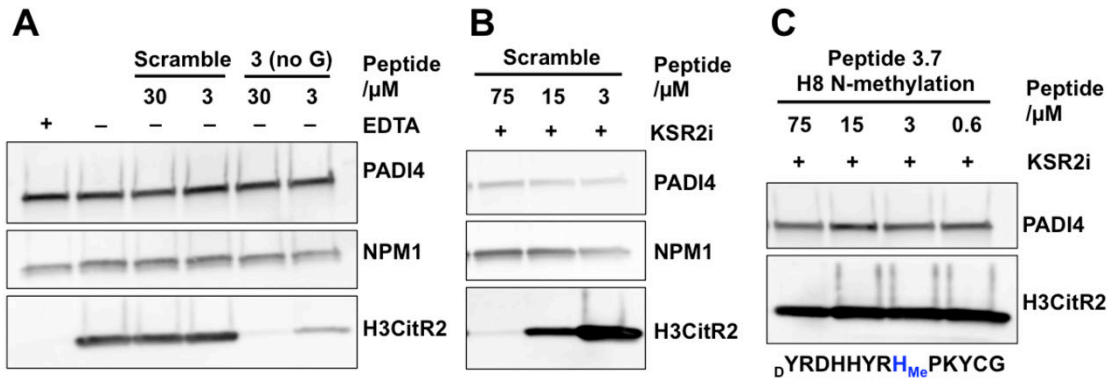


Figure 5.17: Initial testing of control peptides. A randomly shuffled peptide 3 variant was synthesized and tested **A:** *in vitro* in a citrullination lysate assay, compared to Peptide 3 with a single Glycine absent from the sequence (see Figure 5.9D) and **B:** in cells which had been activated using KSR2i treatment for 6 hours. **C:** An alternative peptide 3 variant with a sequence that differs only by backbone N-methylation showed no inhibitory effects in cells. Data in this figure are shown from a single preliminary experiment.

I was unable to finish this work before ending in the lab so this will have to be continued in future work. Nonetheless this quick final experiment emphasizes the need for robust peptide controls. Scrambled versions of activator peptides 11 and 14 as well the pulldown peptide 7 were prepared for this purpose and will be the first place to start. Peptide 3.7 already should provide a good alternative negative control for the inhibitory peptide 3 and peptide 9 provides a control for the activator peptides in the meantime. These experiments will be important for the validation of the reagents described in this chapter and I hope this work will be followed up in the future.

5.9 Discussion

This chapter presents work to develop and characterize a toolkit of cyclic peptide reagents that were shown to inhibit, activate and pull down human PADI4 with efficacy in vitro and in cells. As my work in the lab finished somewhat abruptly, some obvious follow-up experiments and repeats are left to future work. Obtaining formal IC_{50}s and EC_{50}s of the inhibitors and activators respectively will clearly be useful. The method developed for a high content microscopy chemical genetic screen is set up and ready to be used to obtain cellular IC_{50}s and EC_{50}s for the inhibitors and activators (Figure 4.9E) and will also assist cellular toxicity profiling. This work should include selectivity profiling against the different human and mouse PADI paralogues. Preliminary data suggest the inhibitor peptides at least are likely to be specific to human PADI4, as they are not active on mouse PADI4 (personal communication, Dr Walport) and from the precedent of previous peptides identified using the RaPID system{Passioura:2017cc, Suga:2018ii}.

It will also be an interesting curiosity to test these peptides on the newly identified cyanobacterial homologue from Chapter 3.

More exciting follow-up will include demonstrating target engagement of the peptides in cells. This can be done using a cellular thermal shift assay (CETSA). Work on this was initiated and CETSA temperatures for PADI4 are provided with the protocol in the Appendix. Coupling a CETSA experiment with MS/MS would help to identify any off-target effects. From a structural biology perspective, future experiments such as co-crystalisation of the enzyme with peptide could be used to identify where the activator peptides bind allosterically. Identification of an activating binding region by a cyclic peptide would aid mechanistic understanding of PADI4 activation such as by an endogenous protein (such as the candidate calmodulin that was identified from MS/MS data in Chapter 6). Similarly testing whether activators act in tandem with a candidate regulator to activate PADI4 would give an indication of whether they may operate using the same mechanism.

The real advantage of these tools will be realized by the biological experiments they enable. Having potent reversible inhibitors will clearly be very useful, including the first reversible inhibitors to target the active conformation of PADI4. These can be used and characterized in the other cellular systems set up in Chapter 4 (such as to inhibit NET formation in HL60s or primary human neutrophils). The inhibitory peptides clearly also offer potential given the promise shown by PADI4 inhibition in disease[44]{Lewis:2016ul}. Other derivatives of the peptides (other than biotin) would also be worth investigating such as FAM-conjugated peptides for fluorescent detection, since the biotinylated peptide 7 worked very efficiently for pulldown. The real biological novelty is likely to be derived from the activators. This might firstly include experiments to dissect the precise effects of PADI4 from the pleiotropic NET-inducing inflammatory stimuli used at the moment. In the first instance, it will be possible to interrogate whether PADI4 activation is sufficient to induce NET formation in neutrophils. Secondly, they

will be interesting for the role of PADI4 in pluripotency context. Use of activating peptides could ascertain whether activating PADI4 increases reprogramming efficiency. An immediate biological follow-up would to test whether activating peptides might be able to directly increase Wnt transcriptional output in mouse ES cells (a repeat of Figure 4.13). In summary, the ability to activate PADI4 directly will allow a precise dissection of its role and therefore facilitate exploration of the precise role of PADI4 activation in cells. These reagents therefore present exciting possibilities for future work as a wide range of exciting experiments should now be possible.

A final point for discussion is that several variant selections were also designed for future work depending on the success of these first selections. The first involved a variant of the second 'activator' version of the RaPID selection where a histone substrate, with the substrate arginine residue replaced with Cl-amidine, could be used to more comprehensively block the active site. A second idea was to perform the selection against enzyme that had been preincubated in limiting calcium concentrations, where only half of the calcium sites are occupied, as has been similarly shown in crystal structures of PADI2[36]. This could conceivably trap a stabilised intermediate PADI conformation and also act to activate the enzyme. It was additionally hoped that, if we could develop a method to isolate the endogenously activated PADI4 from studies *in vivo* (such as a post-translationally modified variant of PADI4 or a PADI4 linked to an activating partner), selections (both for inhibitors or activators) could be performed either on the endogenously modified enzyme or endogenously activated conformation. This might enable inhibitors that could specifically tune down the deregulated enzyme activation observed in various pathologies.

5.10 References for Chapter 5

1. Luo, Y., Knuckley, B., Lee, Y.-H., Stallcup, M. R. & Thompson, P. R. A fluoroacetamidine-based inactivator of protein arginine deiminase 4: design, synthesis, and in vitro and in vivo evaluation. *J. Am. Chem. Soc.* **128,** 1092–1093 (2006).
2. Luo, Y. *et al.* Inhibitors and inactivators of protein arginine deiminase 4: functional and structural characterization. *Biochemistry* **45,** 11727–11736 (2006).
3. Lewis, H. D. *et al.* Inhibition of PAD4 activity is sufficient to disrupt mouse and human NET formation. *Nat. Chem. Biol.* **11,** 189–191 (2015).
4. Islam, M. S., Leissing, T. M., Chowdhury, R., Hopkinson, R. J. & Schofield, C. J. 2-Oxoglutarate-Dependent Oxygenases. *Annu. Rev. Biochem.* **87,** 585–620 (2018).
5. Suga, H. Max Bergmann award lecture:A RaPID way to discover bioactive nonstandard peptides assisted by the flexizyme and FIT systems. *Journal of Peptide Science* **24,** e3055 (2018).
6. Knipp, M. & Vasak, M. A colorimetric 96-well microtiter plate assay for the determination of enzymatically formed citrulline. *Anal. Biochem.* **286,** 257–264 (2000).
7. Kearney, P. L. *et al.* Kinetic characterization of protein arginine deiminase 4: A transcriptional corepressor implicated in the onset and progression of rheumatoid arthritis. *Biochemistry* **44,** 10570–10582 (2005).
8. Liao, Y. F., Hsieh, H. C., Liu, G. Y. & Hung, H. C. A continuous spectrophotometric assay method for peptidylarginine deiminase type 4 activity. *Anal. Biochem.* **347,** 176–181 (2005).
9. Knuckley, B., Luo, Y. & Thompson, P. R. Profiling Protein Arginine Deiminase 4 (PAD4): a novel screen to identify PAD4 inhibitors. *Bioorg. Med. Chem.* **16,** 739–745 (2008).
10. Moelants, E. A. V., Van Damme, J. & Proost, P. Detection and Quantification of Citrullinated Chemokines. *PLoS ONE* **6,** e28976 (2011).
11. Jones, J. E. *et al.* Synthesis and Screening of a Haloacetamidine Containing Library To Identify PAD4 Selective Inhibitors. *ACS Chem. Biol.* **7,** 160–165 (2011).
12. Nemmara, V. V. & Thompson, P. R. in *Activity-Based Protein Profiling* **420,** 233–251 (Springer, Cham, 2018).
13. Tejeda, E. J. C. *et al.* Noncovalent Protein Arginine Deiminase (PAD) Inhibitors Are Efficacious in Animal Models of Multiple Sclerosis. *J. Med. Chem.* **60,** 8876–8887 (2017).
14. Roberts, R. W. & Szostak, J. W. RNA-peptide fusions for the in vitro selection of peptides and proteins. *PNAS* **94,** 12297–12302 (1997).
15. Nemoto, N., Miyamoto-Sato, E., Husimi, Y. & Yanagawa, H. In vitro virus: bonding of mRNA bearing puromycin at the 3'-terminal end to the C-terminal end of its encoded protein on the ribosome in vitro. *FEBS Letters* **414,** 405–408 (1997).
16. Jongkees, S. A. K. *et al.* Rapid Discovery of Potent and Selective Glycosidase-Inhibiting De Novo Peptides. *Cell Chem Biol* **24,** 381–390 (2017).
17. Ito, K. *et al.* Artificial human Met agonists based on macrocycle scaffolds. *Nat Commun* **6,** 6373 (2015).
18. Hayashi, Y., Morimoto, J. & Suga, H. In vitro selection of anti-Akt2 thioether-macrocyclic peptides leading to isoform-selective inhibitors. *ACS Chem. Biol.* **7,** 607–613 (2012).
19. Yamagishi, Y. *et al.* Natural Product-Like Macrocyclic N-Methyl-Peptide Inhibitors against a Ubiquitin Ligase Uncovered from a Ribosome-Expressed De Novo Library. *Chemistry & Biology* **18,** 1562–1570 (2011).
20. Morimoto, J., Hayashi, Y. & Suga, H. Discovery of Macrocyclic Peptides Armed with a Mechanism Based Warhead: Isoform Selective Inhibition of Human Deacetylase SIRT2. *Angewandte Chemie* **124,** 3479–3483 (2012).
21. Yamagata, K. *et al.* Structural Basis for Potent Inhibition of SIRT2 Deacetylase by a Macrocyclic Peptide Inducing Dynamic Structural Change. *Structure* **22,** 345–352 (2014).

22. Kawamura, A. *et al.* Highly selective inhibition of histone demethylases by de novo macrocyclic peptides. *Nat Commun* **8,** 14773 (2017).
23. Song, X., Lu, L.-Y., Passioura, T. & Suga, H. Macrocyclic peptide inhibitors for the protein-protein interaction of Zaire Ebola virus protein 24 and karyopherin alpha 5. *Org. Biomol. Chem.* **15,** 5155–5160 (2017).
24. Matsunaga, Y., Bashiruddin, N. K., Kitago, Y., Takagi, J. & Suga, H. Allosteric Inhibition of a Semaphorin 4D Receptor Plexin B1 by a High-Affinity Macrocyclic Peptide. *Cell Chem Biol* **23,** 1341–1350 (2016).
25. Vinogradov, A. A., Yin, Y. & Suga, H. Macrocyclic Peptides as Drug Candidates: Recent Progress and Remaining Challenges. *J. Am. Chem. Soc.* **141,** 4167–4181 (2019).
26. Letschert, K., Faulstich, H., Keller, D. & Keppler, D. Molecular characterization and inhibition of amanitin uptake into human hepatocytes. *Toxicol. Sci.* **91,** 140–149 (2006).
27. McLellan, N. L. & Manderville, R. A. Toxic mechanisms of microcystins in mammals. *Toxicol Res (Camb)* **6,** 391–405 (2017).
28. Vives, E., Brodin, P. & Lebleu, B. A truncated HIV-1 Tat protein basic domain rapidly translocates through the plasma membrane and accumulates in the cell nucleus. *J. Biol. Chem.* **272,** 16010–16017 (1997).
29. Derossi, D., Joliot, A. H., Chassaing, G. & Prochiantz, A. The third helix of the Antennapedia homeodomain translocates through biological membranes. *J. Biol. Chem.* **269,** 10444–10450 (1994).
30. Schmidt, N., Mishra, A., Lai, G. H. & Wong, G. C. L. Arginine rich cell penetrating peptides. *FEBS Letters* **584,** 1806–1813 (2010).
31. Futaki, S., Hirose, H. & Nakase, I. Arginine-rich peptides: methods of translocation through biological membranes. *Curr Pharm Des* **19,** 2863–2868 (2013).
32. Zaro, J. L. & Shen, W.-C. Cationic and amphipathic cell-penetrating peptides (CPPs): Their structures and in vivo studies in drug delivery. *Front. Chem. Sci. Eng.* **9,** 407–427 (2015).
33. Walport, L. J., Obexer, R. & Suga, H. Strategies for transitioning macrocyclic peptides to cell-permeable drug leads. *Curr. Opin. Biotechnol.* **48,** 242–250 (2017).
34. Abramson, A. *et al.* An ingestible self-orienting system for oral delivery of macromolecules. *Science* **363,** 611–615 (2019).
35. Arita, K. *et al.* Structural basis for Ca2+-induced activation of human PAD4. *Nature Structural & Molecular Biology* **11,** 777–783 (2004).
36. Slade, D. J. *et al.* Protein arginine deiminase 2 binds calcium in an ordered fashion: implications for inhibitor design. *ACS Chem. Biol.* **10,** 1043–1053 (2015).
37. Darrah, E., Rosen, A., Giles, J. T. & Andrade, F. Peptidylarginine deiminase 2, 3 and 4 have distinct specificities against cellular substrates: novel insights into autoantigen selection in rheumatoid arthritis. *Ann Rheum Dis* **71,** 92–98 (2012).
38. Christophorou, M. A. *et al.* Citrullination regulates pluripotency and histone H1 binding to chromatin. *Nature* **507,** 104–108 (2014).
39. Ying, Q.-L. *et al.* The ground state of embryonic stem cell self-renewal. *Nature* **453,** 519–523 (2008).
40. Theunissen, T. W. *et al.* Nanog Overcomes Reprogramming Barriers and Induces Pluripotency in Minimal Conditions. *Current Biology* **21,** 65–71 (2011).
41. Rogers, J. M., Passioura, T. & Suga, H. Nonproteinogenic deep mutational scanning of linear and cyclic peptides. *PNAS* **115,** 10959–10964 (2018).
42. Darrah, E. *et al.* Erosive Rheumatoid Arthritis Is Associated with Antibodies That Activate PAD4 by Increasing Calcium Sensitivity. *Science translational medicine* **5,** 186ra65–186ra65 (2013).
43. Slack, J. L., Causey, C. P., Luo, Y. & Thompson, P. R. Development and Use of Clickable Activity Based Protein Profiling Agents for Protein Arginine Deiminase 4. *ACS Chem. Biol.* **6,** 466–476 (2011).
44. Lewis, H. D. & Nacht, M. iPAD or PADi—'tablets' with therapeutic disease potential? *Current Opinion in Chemical Biology* **33,** 169–178 (2016).

Chapter 6: Towards the Physiological Activation of PADI4

6.1 Introduction

In Chapter 4, cellular systems were set-up in which PADI4 could be physiologically activated. The aim for Chapter 6 was to use proteomic analysis (affinity purification tandem mass spectrometry, AP-MS/MS) to look firstly for any PTMs on PADI4 protein and secondly for any interacting proteins that differ between resting and activated PADI4. This would be used to collate a list of putative regulatory candidates for the endogenous activation of PADI4 protein. Finally one potential regulatory candidate, calmodulin, was taken forward for validation and explored in biochemical assays where it activated PADI4 in vitro.

6.1.1 Past attempts to do MS/MS on PADI4

Approaches to analyze the PADI4 interactome and PTMs on PADI4 have been attempted elsewhere, including to look for regulatory events[1-6]. The hope for this work was that the new tractable activation conditions in cellular systems and a carefully optimized MS/MS approach established in this Chapter might result in an improved chance of finding regulatory molecular events from previous attempts in the literature. Any candidates that differ across cellular treatments from the MS/MS analyses would need to be carefully validated for their capacity to activate PADI4 *in vitro* and in cells.

6.1.2 Objectives

- Perform proteomic analysis of PADI4 to identify candidates for the physiological regulation of the enzyme in cells.
- Validate a chosen candidate for the potential to activate PADI4.

6.2 Establishing methods for isolation and MS/MS analysis of PADI4

Section 6.2 describes the optimization of conditions to isolate human PADI4 from cells and the set-up of computational pipelines to process and interpret the raw MS/MS data.

6.2.1 Optimizing lysis conditions

I started by setting up lysis conditions to efficiently extract PADI4 from cells. The best conditions for preserving intracellular protein-protein interactions are clearly very different from those required to preserve and detect direct PTMs to the bait protein. In the first instance, it is desirable to preserve interactions using a gentle lysis. In the second instance, provided that both the pulldown is still efficient and the desired PTMs are not labile after stringent lysis, there is an advantage to aggressive lysis procedures as more protein can be isolated and enzymes present in the cell can be rapidly inactivated prior to isolation.

Two different lysis protocols were established for isolating PADI4 from cells for these purposes. The first was designed as a compromise between being able to disrupt the nuclear membrane (and release nuclear PADI4) while at the same time minimizing the disruption of intracellular protein-protein interactions (0.5% NP-40 and benzonase, Chapter 2.3). A second protocol was established to rapidly lyse the cell and also rapidly denature any kinases, phosphatases or other enzymes present in the lysate (1% SDS, with boiling). The lysate is diluted tenfold to reduce the SDS concentration for efficient pulldown. It was hoped this second protocol would stabilize the state of direct PTMs present on PADI4. In both instances a cocktail of kinase and phosphatase inhibitors was added to preserve the endogenous phosphorylation state of proteins in the lysates.

6.2.2 Isolation of PADI4 from cells

Following a series of optimization steps, I was able to isolate PADI4 from HL-60 cells and PADI4-stable mES cells using immunoprecipitation with a commercially available antibody (Chapter 2.3, Figure 6.1). Importantly, this was done while disrupting and retaining the chromatin fraction of the cell, as a portion of PADI4 is expected to be bound to chromatin[4]. In order to reduce the extensive signal from the heavy chain of the antibody observed in Western blotting (that runs close to the molecular weight of human PADI4),

the crosslinking agent BS^3 was used to conjugate the primary antibody to the Protein A/Protein G beads. Pulldown of PADI4 using peptide 7 was also achieved and is covered in the previous chapter (Section 5.7).

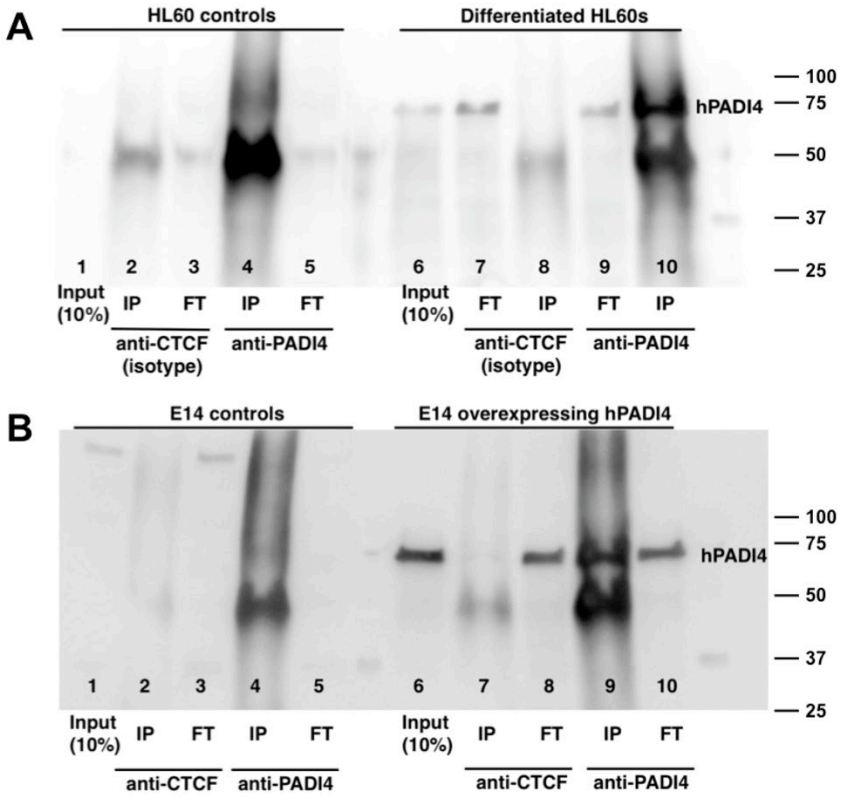


Figure 6.1: Immunoprecipitation using anti-human PADI4 antibody ab50332. **A:** Immunoblot analysis of PADI4 immunoprecipitated from undifferentiated HL-60 cells where human PADI4 is not expressed (Lanes 1-5) and from terminally differentiated HL-60 cells, where human PADI4 is expressed, using anti-PADI4 and an isotype control anti-CTCF antibody (Lanes 6-10). The lower molecular weight band corresponds to the denatured heavy chain of the PADI4 antibody. The band labeled as PADI4 appears at the expected molecular weight for PADI4, which can be detected in the input (Lane 6). **B:** Immunoblot analysis of PADI4 immunoprecipitated from control-stable cells (Lanes 1-5) and from PADI4-stable cells, using anti-human PADI4 and an isotype control anti-CTCF antibody (Lanes 6-10). The lower molecular weight band corresponds to the denatured heavy chain of the PADI4 antibody. The band labeled as PADI4 appears at the expected molecular weight for PADI4, which can be detected in the input (Lane 6). Abbreviations IP = contents of immunoprecipitation, FT = flow through. Data shown are representative of n = 2.

6.2.3 Mass spectrometry data analysis approaches

Two different software approaches were used for raw mass spectrometry data analysis. Although MaxQuant is used widely by the MRC IGMM Institute Mass Spectrometry service, it presents a number of challenges for the purposes in this Chapter. The sophisticated label free protein quantification (LFQ) employed in MaxQuant firstly assumes broad proteome coverage and secondly that most proteins will be unchanging across different samples. This is less applicable, however, to affinity pulldowns where the proteins detected after isolation will be inherently biased by design: the point of the experiment is specifically to identify those proteins enriched by the bait, and not to obtain broad proteome coverage. The second major problem is that searching for variable post-translational modifications is hugely computationally expensive, particularly for multiple modifications as additional variable PTM searches scale exponentially in computational time. In this case, we had hoped to map as comprehensive a list of PTMs as possible to PADI4 protein. MaxQuant was therefore only used for generating initial interactome datasets. It performs very effective feature matching and identification across different runs. Where presented in this Chapter, the MaxQuant LFQ intensity for a given protein was normalized to the LFQ intensity for trypsin. Samples were then averaged and the $\log_2$(ratio) calculated for a given condition.

I undertook a second approach to MS/MS analysis using a computational pipeline with PEAKS7.0. PEAKS software performs complete *de novo* peptide sequencing of the data *in silico* before doing database searches. This involves label free quantification and statistical analysis but does so across replicates that are assigned to be within a single group of the same treatment. This is a particularly useful way to handle exploration of the differential interactome under different cellular stimuli. Our hypothesis is that a limited number of proteins will change significantly across samples treated differently in this instance, but that replicates by contrast ought to be relatively unchanged. The data were then presented as heat maps of

log_2ratio of enrichment and depletion as well as assigning and ranking any proteins that significantly differ between cellular treatments. These are provided in the order of their $-10log_{10}$(p-value) where a value >20 corresponds to a p-value < 0.01.

A more detailed account of the calculation of significance is given below. Briefly, the LFQ analysis by PEAKS is based on having multiple samples per group (biological condition). Feature detection is performed separately on each sample and LFQ is then based on the relative intensities of peptide features detected across multiple samples. Features are aligned based on retention time alignment and an Expectation-Maximization algorithm means that more overlapped features can be detected. Normalization was performed against the total ion count of the samples. Peptide feature significance is based on a two-tailed p-value with respect to deviation of the feature vector from a lognormal distribution corresponding to both the largest group ratio (features which change most across different conditions) and the peptide quantification quality (best detected peptides). Protein significance is then inferred from the supporting peptide feature significances. These are weighted by the intensity rank of each peptide feature corresponding to that protein as well as the correlation between the relative abundance of the protein and the relative abundance of its supporting peptides (further details can be found in the PEAKS documentation).

The second significant advantage of the PEAKS computational pipeline is the *de novo* peptide sequencing approach. This means all modified peptides are sequenced and identified *a priori* which allows any combination of post-translational modifications to be identified afterwards. This bypasses the prohibitively computationally expensive problems associated with combined variable modifications that are encountered in MaxQuant.

6.2.4 Coverage of protein for PTMs

One of the major challenges in obtaining comprehensive PTM analysis of a single target protein is protein coverage. MS/MS still relies on proteolytic enzymatic digestion to produce peptide fragments of a certain size that can be readily detected by MS/MS (this is generally taken to be a limit of 6-30 amino acids). For the most part, the MS/MS workhorse enzyme is trypsin, which cuts at lysine and arginine residues. If these residues are either too frequently or too sparsely distributed in the protein sequence of interest, detectable tryptic fragments will not be generated within the appropriate window, which almost invariably results in incomplete protein coverage. I performed multiple digestions using various enzymes *in silico*. Assuming 100% digestion efficiency, I found that the theoretical coverage of PADI4 was 61.1% with single digestion using Trypsin, but could be improved up to 97.9% with the use of four enzymes for digestion (Trypsin, Chymotrypsin, AspN, and GluC)[7]. The computational pipeline in PEAKS is designed for use in experiments that include multiple digestions in the future.

6.3 Overview of experiments to analyze PADI4 by MS/MS

The MS/MS experiments to analyze PADI4 that were performed are outlined in table below (Figure 6.2) and referred to throughout in the text by the name given in the first column, but are introduced again in each section for clarity of reading. Section 6.4 summarizes the MS/MS attempts to analyze PADI4 for regulatory PTMs. Section 6.5 summarizes experiments to identify the differential interactome of PADI4 between resting and activated cells. Section 6.6 then describes experimental validation of one of the candidates derived from Section 6.5 (calmodulin) and biochemical assays to validate its effects on PADI4 activation.

Name	Cell Line	Methods	Cell treatments	Pull-down	Figure
EXP1	HL60	PTMs Interactome	1) Undifferentiated HL60s 2) Differentiated HL60s 3) Differentiated HL60s + Calcium ionophore	Antibody	6.3 6.5 6.6
EXP2	HL60	PTMs RIME	1) Undifferentiated HL60s 2) Differentiated HL60s 3) Differentiated HL60s + LPS 4) Differentiated HL60s + LPS + Calcium ionophore	Antibody	6.7
EXP3	mES	PTMs RIME	1) mES control 2) mES PADI4 stable 3) mES PADI4 stable + Calcium ionophore	Antibody	6.7
EXP4	mES	Interactome	1) mES control 2) mES PADI4	Peptide_7	5.16
EXP5	mES	PTMs Interactome	1) mES PADI4 2) mES PADI4 + CHIR99021 3μM 3) mES PADI4 + Cl-amidine 100μM 4) mES PADI4 + Cl-amidine 100μM + CHIR99021 3μM	Antibody Peptide_7	6.8 6.10 6.13

Figure 6.2: Summary of MS/MS experiments conducted in this chapter.

6.4 Analysis of PTMs identified on PADI4

6.4.1 Analysis of PTMs identified on PADI4 in EXP1

As a first analysis, MS/MS was conducted on a published PADI4 activation condition in the promyeloblast HL-60 cancer cell line[8]. In EXP1, human PADI4 was immunoprecipitated from undifferentiated HL-60 cells (where PADI4 is not expressed), differentiated HL60 cells (where PADI4 expression is induced, but no activity can be detected) and from differentiated HL-60 cells treated with calcium ionophore (where PADI4 is present and has been activated) (Figure 4.2B and Figure 4.4).

PTMs to PADI4 in EXP1 were analyzed using the computational pipeline in PEAKS7.0 (Section 6.2.3). The coverage obtained for the PADI4 sequence after this initial experiment was 49% after tryptic digestion with the distribution of coverage across the PADI4 sequence shown in Figure 6.3 (compared to theoretical maximum coverage of 61.1%). No putative regulatory PTMs were detected on PADI4 from EXP1. One surprising finding, however, was that two peptides containing the active site cysteine in PADI4 were found to specifically modified in the resting condition, with evidence of this modification specifically depleted from the activated condition (Figure 6.3, $\log_2$ratio < –3.5). The modification detected, carbamidomethylation, is an artificial modification introduced during the mass spectrometry workflow in which cysteine residues are converted to carbamidomethylated cysteines prior to MS/MS. Interestingly, if a modification were endogenously present at cysteine, then the efficiency of introduced carbamidomethylation in the mass spectrometry workup would differ. This serendipitous observation is a similar approach to the method specifically designed to detect cysteine nitrosylation in the literature. Cys S-nitrosylation modifications are otherwise labile under typical mass spectrometry conditions and so cannot usually be detected. If the active site were differentially protected by an endogenous PTM such as cys S-nitrosylation, this finding could potentially be very interesting.

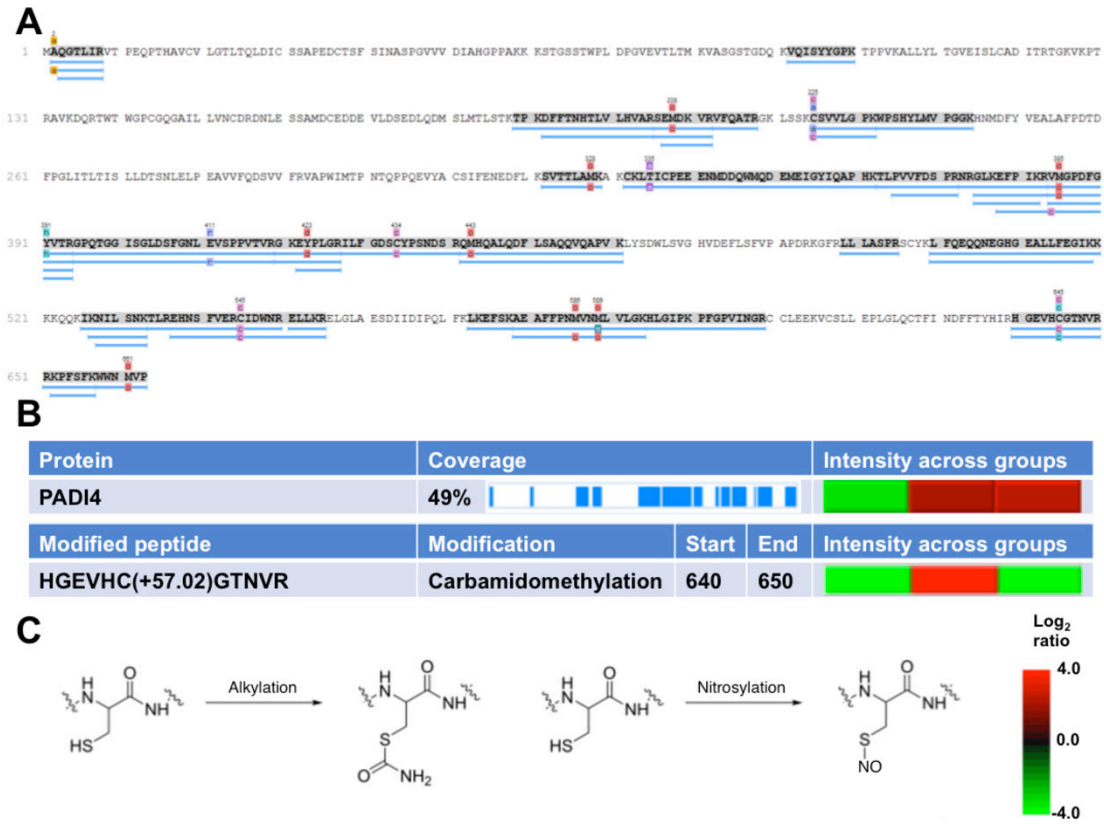


Figure 6.3: PTM analysis of PADI4 analyzed from EXP1. **A:** Output of all PTMs detected to PADI4 in EXP1. Four replicates were performed per condition. 49% of the protein was covered. Blue lines below the PADI sequence correspond to identified peptide features. **B:** Differential carbamidomethylation state of active site cysteine 645 of PADI. Modified peptide is enriched in the condition where PADI4 was isolated from differentiated HL60s (resting), but depleted from differentiated HL60s that had been treated with calcium ionophore (activated). **C:** Schematic showing possible hypothesis that the active cysteine may be in a different endogenous state in the inactive form. This could be due to a different redox state and could be endogenously nitrosylated.

PTMs on other proteins were then used as an estimate of the success of PTM detection in this experiment overall given the lack of PTMs identified on the target protein, PADI4. To do this, I analyzed the known PADI4 substrate Histone 1.2[4]. Citrullination of Arginine 54, located within the DNA binding domain of H1, is particularly interesting as it was shown to result in its displacement from chromatin and in global chromatin decondensation[4]. This protein was co-enriched in both resting and activated PADI4 conditions, but not in the control condition where PADI4 is absent (Figure 6.4) suggesting it has derived from co-IP. Coverage of histone 1.2 was found to be 73% and

multiple modified peptides could be detected including 17 phosphorylated, 33 acetylated, 21 hydroxylated and 14 citrullinated features. This included detection of phosphorylations to Thr4, Ser36, Tyr71 and Ser78; acetylation to Lys17, Lys21, Lys27, Lys75 and Lys90 and citrullinations to Arg33 and Arg54 (Figure 6.4). Citrullinated peptides (with a 0.98Da mass shift) containing Histone 1 Arginine 54 (a known target of PADI4) were identified as being enriched specifically in the activated PADI4 condition, but were depleted from the resting condition (Figure 6.4).

It is therefore pleasing from this experiment that, in the binary case of citrullination of Arg 54 of histone 1.2, a known PTM could be differentially identified from the PEAKS computational pipeline. In addition, phosphorylations were detectable such as those to histone H1.2. For less binary modifications, however, observing different levels of a modified peptide against the unmodified form will be challenging as these peptides will behave differently (and non-quantitatively with respect to each other) in the mass spectrometer.

Figure 6.4: PTM analysis of Histone H1.2 (P16403) that was identified by co-IP of human PADI4 from EXP1. Blue lines below the PADI sequence correspond to identified peptide features. 73% of the protein was covered. Four replicates were performed per condition. **A:** Output of all PTMs detected to the H1.2 detected in EXP1. Multiple modified

peptide features could be detected including 17 phosphorylated, 33 acetylated, 21 hydroxylated and 14 citrullinated peptide features. This included phosphorylated Thr4, Ser36, Tyr71, Ser78; acetylated Lys17, Lys21, Lys27, Lys75, Lys90 and citrullinated Arg33 and Arg54. **B:** The differential citrullination state of known modified arginine 54 is evidenced by two different modified peptide features that were found to be enriched only in the sample isolated from differentiated HL60s treated with calcium ionophore (with activated PADI4).

6.4.2 Analysis of PTMs identified on PADI4 in EXP5

MS/MS was then conducted on the new PADI4 activating stimuli identified in Chapter 4 to look for any PTMs (Section 4.2.3.3). In this experiment (EXP5), human PADI4 was isolated from 1) PADI4-stable cells (treated for 45 minutes with DMSO) and 2) PADI4-stable cells that were treated with 3μM CHIR99021 (GSK3β inhibitor) for 45 minutes. Samples were taken to confirm successful activation of human PADI4 occurred only in the treated samples by immunoblotting (Figure 4.9D). Samples were prepared for isolation by immunoprecipitation with a commercial antibody to human PADI4 (3 replicates per treatment), and by pulldown using biotinylated peptide 7 (Chapter 5) (3 replicates per treatment) for a total of 6 replicates per treatment. Analysis was performed using the PEAKS computational pipeline and the data from the peptide pulldown and antibody were pooled and all identified PTMs were collected (Figure 6.5).

The PTM data are presented on the sequence of PADI4 with blue bars beneath showing the peptide features identified (Figure 6.5). The observed coverage was increased to 63% of the protein (417/663 amino acids), which is slightly higher than the theoretical maximum coverage assuming 100% digestion efficiency (61.1%). This is possible since incomplete tryptic digestion can increase the MS/MS coverage of smaller peptides. 25 PTMs were identified including a phosphorylation at Ser433 and a ubiquitination at Lys525 (Figure 6.5). As the cellular activation of PADI4 occurs downstream of single kinase inhibition (GSK3β), the direct phosphorylation to PADI4 is particularly interesting. Both unphosphorylated and phosphorylated peptide features were detected indicating it is sub-stoichiometric and therefore

potentially regulatory as it may differ between different conditions. GSK3β typically has a priming phosphorylation site located three amino acid residues away from the target Serine/Threonine and is a Proline-directed kinase. Intriguingly this phosphoSer433 has a second Ser site located three residues C-terminally. In addition this downstream Ser436 is adjacent to a Pro residue so this site could be a direct target of GSK3β. These serines are conserved in PADI2 and the loop containing these two Serines is located close to the active site (Figure 6.5).

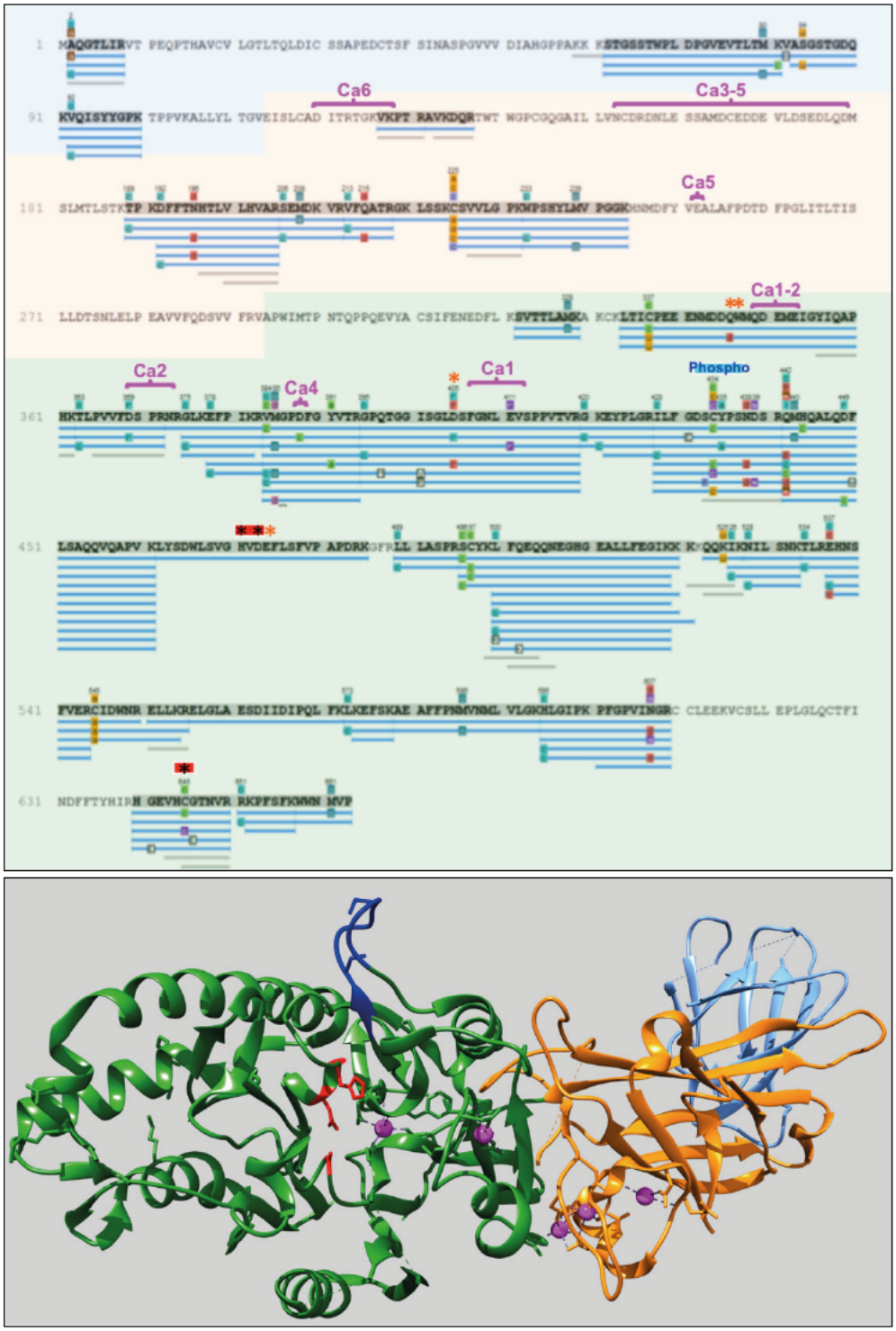


Figure 6.5: PTM analysis of PADI4 analyzed from EXP5. Top panel: Output of all PTMs detected to PADI4 in EXP1. 63% of the protein was covered. Blue lines below the PADI sequence correspond to identified peptide features. 64 peptides were identified with 25 PTMs, including K525 Ubiquitination, and S433 Phosphorylation. The PAD_N domain is shaded in blue, the PAD_M domain in orange and the PAD_C domain in green. Magenta brackets above the sequence denote calcium binding site regions as they align to PADI2;

noting that calcium binding site 6 is not conserved in PADI4. Red boxes denote the location of the catalytic triad with residues given by a black asterisk. An orange asterisk is included above other active site resides. A cyan blue box denotes the location of the phosphorylated motif identified by MS/MS experiments in this chapter. Data were collated from 12 replicates; six replicates were used per condition. **Bottom panel:** Crystal structure of calcium bound PADI4 (1wd9). The PAD_N domain is coloured in cornflower blue, the PAD_M domain in orange and the PAD_C domain in green. Calcium ions are coloured in magenta, active site residues in red, and the phosphorylated motif is shown in blue: D**S**CYP**S**NDS.

6.5 Identifying PADI4 interacting proteins differing between resting and activated conditions

Experiments were then undertaken to identify any proteins that interact differently with resting or activated PADI4 and may thereby modulate enzyme activation.

6.5.1 Differential interactome analysis of EXP1

Similarly to the PTM analysis, the first attempt (EXP1) was conducted on the promyeloblast HL-60 cancer cell line using an established PADI4 activation condition (Figure 6.6). Group 1 refers to the first condition where human PADI4 was immunoprecipitated from undifferentiated HL-60 cells (where PADI4 is not expressed), Group 2 to differentiated HL60 cells (where PADI4 expression is induced, but no activity can be detected) and Group 3 to differentiated HL-60 cells treated with calcium ionophore (where PADI4 is present and has been activated) (Figure 4.2B and 4.4).

After processing the raw data, the top proteins that were identified in PEAKS analysis were clustered and $\log_2$ratio enrichment or depletion was visualized across the three conditions in EXP1 (Figure 6.6). As expected, PADI4 protein (Q9UM07) is enriched in the two conditions using differentiated HL-60 cells, and absent from undifferentiated HL-60s, where it is not expressed and is a useful sense check to show the AP-MS/MS approach is working as expected.

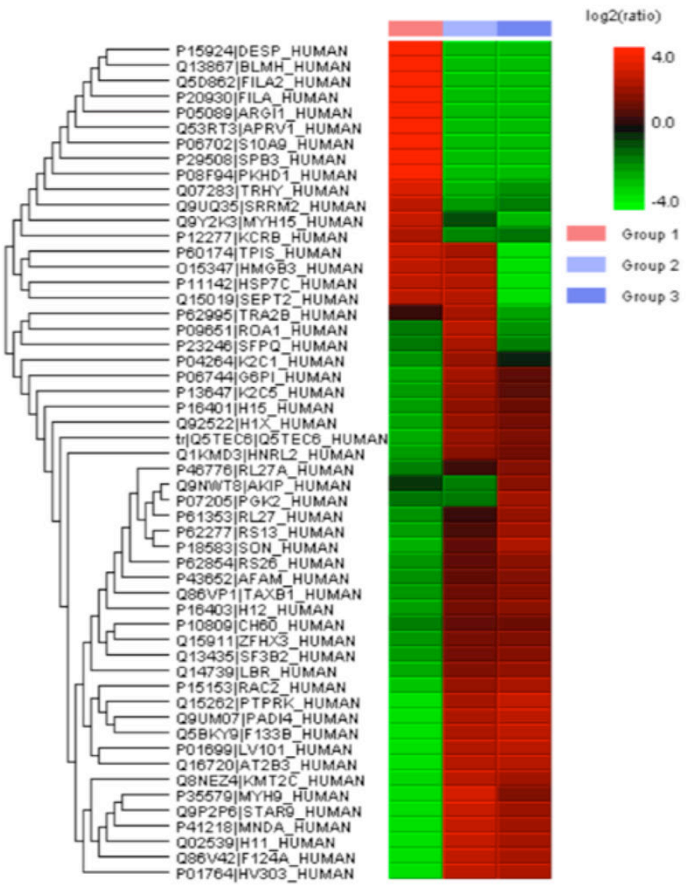


Figure 6.6: Mass spectrometry dataset from EXP1 with clustering analysis performed using Peaks7.0 software. Group numbers correspond to treatment numbers in Figure 6.2 – group 1 are undifferentiated HL60 cells (no PADI4 expression), group 2 are differentiated HL60 cells (inactive PADI4), and Group 3 are differentiated HL60 cells treated with calcium ionophore (activated PADI4). Four replicates were performed per condition.

Many of the top interacting proteins that differ between samples and are enriched in the two PADI4-containing conditions are known substrates of PADI4 (Figure 6.6). The list includes Histone 1 variants, H3 variants, Protein SON and HNRNPs (Figure 6.6). This is good validation that the method is identifying real interactors. To analyze this systematically, I then compared the top proteins from this MS/MS experiment to an unpublished mass spectrometry dataset obtained by Maria Christophorou and Prof Nielsen that represents an extensive list of citrullinated proteins under the same activating conditions used in my experiment. A majority of the top identified interacting proteins were also found to be substrates in this so-called 'citrullinome' dataset (23 out of 37) (Figure 6.7). The overlap between these two sets (or

the proportion of citrullinated proteins identified) is much greater than is expected by chance – given the citrullinome dataset contains 3266 out of an assumed total of ~20000 proteins (the representation factor = 3.8, $p < 4.3 \times 10^{-10}$, modelled using the hypergeometric probability distribution). The expected number of expected substrates is six (total in set 1 x total in set 2 divided by the overall number of proteins), and the greatest expected by chance is an overlap of 10 substrates (greatest number giving a p value > 0.05 under the hypergeometric probability distribution).

Enriched over control	Name	#of Cit sites identified in MAC set	1	2	3
TRA2B	Transformer-2 protein homolog beta	6			
HNRNPA1	Heterogeneous nuclear ribonucleoprotein A1	9			
SFPQ	Splicing factor, proline- and glutamine-rich	6			
KRT1	Keratin, type II cytoskeletal 1	–			
G6PI	Glucose-6-phosphate isomerase	4			
K2C5	Keratin, type II cytoskeletal 5	–			
HIST1H1B	Histone H1.5	1			
H1FX	Histone H1x	10			
HIST2H3PS2	Histone H3	2			
HNRNPUL2	Heterogeneous NRP U-like protein 2	14			
RPL27A	60S ribosomal protein L27a	8			
AURKAIP1	Aurora kinase A-interacting protein	–			
PGK2	Phosphoglycerate kinase 2	–			
RPL27	60S ribosomal protein L27	2			
RPS13	40S ribosomal protein S13	2			
SON	Protein SON	6			
RPS26	40S ribosomal protein S26	9			
AFAM	Afamin	–			
TAX1BP1	Tax1-binding protein 1	3			
HIST1H1C	Histone H1.2	1			
HSPD1	60 kDa heat shock protein, mitochondrial	5			
ZFHX3	Zinc finger homeobox protein 3	–			
SF3B2	Splicing factor 3B subunit 2	18			
LBR	Lamin B receptor	11			
RAC2	Ras-related C3 botulinum toxin substrate 2	3			
PTPRK	Receptor-type tyrosine-protein phosphatase k	–			
PADI4	Peptidyl arginine deiminase 4	4			
F133B	Protein FAM133B	–			
IGLV1-44	Immunoglobulin lambda variable 1-44	–			
AT2B3	Plasma membrane Ca^{2+}-transporting ATPase 3	–			
KMT2C	Histone-lysine N-methyltransferase 2C	–			
MYH9	Myosin-9	36			
STAR9	StAR-related lipid transfer protein 9	–			
MNDA	Myeloid cell nuclear differentiation antigen	4			
HIST1H1A	Histone H1.1,	1			
FAM124A	Protein FAM124A	–			
IGHV3-23	Immunoglobulin heavy variable 3-23	–			

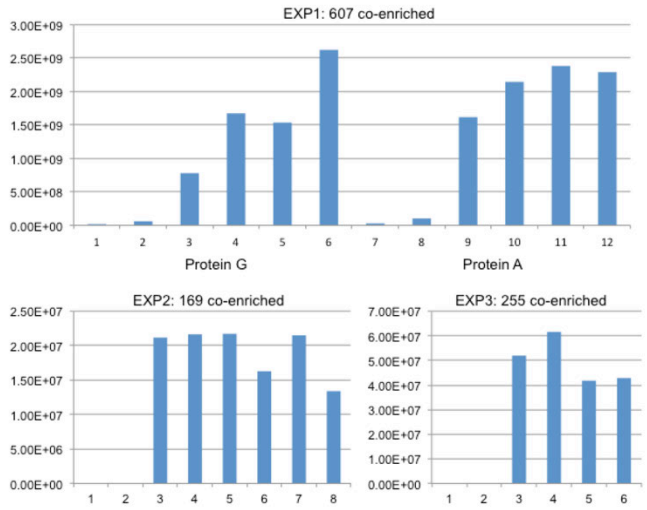
Figure 6.7: Mass spectrometry data from EXP1 with analysis performed using Peaks7.0 software. PADI4 was immunoprecipitated from non-differentiated HL-60 cells as a

control (no PADI4 expression), from differentiated HL-60 cells in a resting condition (where PADI4 protein has been induced), and in activated HL-60 cells after A23187 treatment. Group numbers correspond to the treatment numbers in Figure 6.2.The top proteins with differential detection between control and PADI4 containing conditions are shown in the table. These proteins were compared to a dataset obtained by Maria Christophorou (labelled MAC set in the table), which identified citrullination substrates in the same cell-types and with the same activation condition. Many of the candidates identified as PADI4 interacting proteins are also citrullinated in the other dataset. The number of independently identified citrullination sites identified for these proteins is given in column 3. Log_2 enrichment of the protein in each condition is displayed as a heatmap in column 4. Data were collated from four replicates per condition.

Given the data from Chapter 5 that showed peptides activating allosterically, it is probable, but by no means conclusively true, that *bona fide* PADI4 substrates are a less likely source of candidate regulators. By this logic, two possibly interesting non-substrate proteins appeared to be exclusively enriched in the activating conditions: PGK2 and AURKAIP. PGK2 is a testis-specific enzyme found in mitochondrial metabolism and, at least in the opinion of this author, therefore unlikely to be interesting with respect to PADI4 regulation; but AURKAIP, although partly pulled down in the undifferentiated HL60 cells, may be an interesting candidate to validate further. On the other hand, it is of course also possible that proteins that interact may also be modified if in close proximity to the enzyme. Ruling out substrates *a priori* therefore does not seem to be a good idea with respect to this particular dataset. Candidates identified here are not discussed further for reasons of space, but the list of candidates with enrichments is provided (Figure 6.7). It is notable that the full list of proteins identified as interacting with PADI4 that was obtained with MaxQuant is extensive, showing very efficient AP-MS/MS but also in some ways shows the inherent difficulty in prioritizing candidates without stringent criteria as it is not possible to know when MS/MS analysis might be saturating.

6.5.2 Differential interactome analysis of EXP2 and EXP3: transient PADI4 interactors

Primarily to see if it was possible to identify more transient interactors, a crosslinking Mass spectrometry approach was then attempted on cell conditions established in Chapter 4. The Rapid immunoprecipitation mass spectrometry of endogenous proteins (RIME) protocol was used, which makes use of a formaldehyde crosslinking step before lysis, similar to a chromatin immunoprecipitation (ChIP) protocol[9]. The RIME protocol was performed on two sets of cell treatments (EXP2 from HL-60 cells and EXP3 from mouse ES cells) and datasets were analyzed (Figure 6.8). In EXP2, human PADI4 was immunoprecipitated from undifferentiated HL-60 cells (where PADI4 is not expressed), from differentiated HL60 cells (where PADI4 expression is induced, but no activity can be detected), from differentiated HL60 cells primed with LPS, and from differentiated HL60 cells primed with LPS and treated with calcium ionophore (as in Figure 4.4B). In EXP3, human PADI4 was immunoprecipitated from control-stable cells (mouse ES cells from which no human PADI4 should be isolated), PADI4-stable cells, and from PADI4-stable cells treated with calcium ionophore for six hours (as in Figure 4.6A). In these experiments, cells were crosslinked with formaldehyde before lysis, pulldown and MS/MS analysis.



Gene	UniProt	Gene name
H1FX	Q92522	Histone H1x
LBR	Q14739	Delta(14)-sterol reductase/Lamin B receptor
PROC	P04070	Osteocalcin/Vitamin K-dependent protein C
PDAP1	Q13442	28 kDa heat- and acid-stable phosphoprotein
HMGN4	O00479	High mobility group nucleosome-binding domain-containing protein 4
LILRB3	O75022	Leukocyte immunoglobulin-like receptor subfamily B member 3
SLIT2	O94813	Slit homolog 2 protein
LRRC69	Q6ZNQ3	Leucine-rich repeat-containing protein 69
KRT2	P35908	Keratin, type II cytoskeletal 2 epidermal
KNG1	P01042	Kininogen-1
HMGB1	P09429	High mobility group protein B1
H3F3C	Q6NXT2	Histone H3.3C
HIST1H4A	P62805	Histone H4
RPL27A	P46776	60S ribosomal protein L27a
H2AC11	P0C0S8	Histone H2A type 1
HIST1H1E	P10412	Histone H1.4

Gene	UniProt	Gene name
Q5SSE9	ABCA13	ATP-binding cassette sub-family A member 13
Q8C6P8	Zfp57	Zinc finger protein 57
P15864	Hist1h1c	Histone H1.2
P14602	Hspb1	Heat shock protein beta-1
P12382	Pfkl	ATP-dependent 6-phosphofructokinase, liver type
O08583	Alyref/ THOC4	Alyref export factor/THO complex subunit 4
Q4V9W2	Srek1ip1	Protein SREK1IP1
P09411	Pgk1	Phosphoglycerate kinase 1
Q8VH51	Rbm39	RNA-binding protein 39
Q3UL36	Arglu1	Arginine and glutamate-rich protein 1
Q6NV83	U2surp	U2 snRNP-associated SURP motif-containing protein
Q3UQU0	Brd9	Bromodomain-containing protein 9
P63038	Hspd1	60 kDa heat shock protein, mitochondrial
Q99020	Hnrnpab	Heterogeneous nuclear ribonucleoprotein A/B
Q60668	Hnrnpd	Heterogeneous nuclear ribonucleoprotein D0
P57776	Eef1d	Elongation factor 1-delta
P49312	Hnrnpa1	Heterogeneous nuclear ribonucleoprotein A1
Q9Z315	Sart1	U4/U6.U5 tri-snRNP-associated protein 1
P62751	Rpl23a	60S ribosomal protein L23a
P47915	Rpl29	60S ribosomal protein L29
Q9CX86	Hnrnpa0	Heterogeneous nuclear ribonucleoprotein A0
Q9CY58	Serbp1	Plasminogen activator inhibitor 1 RNA-binding protein
P47911	Rpl6	60S ribosomal protein L6

Figure 6.8: Comparison of LFQ intensities for PADI4 across different trial MS/MS experiments. A: The LFQ intensity for PADI4 is presented on the y-axis. For EXP1, there are duplicate conditions for Protein G (Lanes 1-6), and Protein A (Lanes 7-12) to capture the antibody. **B:** Table of human proteins that were at least threefold co-enriched across all PADI4-isolated conditions in EXP2 after normalization against trypsin (from HL-60 cells). **C:** Table of proteins that were at least threefold co-enriched across all PADI4-isolated conditions in EXP3 after normalization against trypsin (from mouse ES cells). In EXP1, PADI4 was immunoprecipitated from 1) non-differentiated HL-60 cells as a control (no PADI4 expression) (Bars 1-2 and 7-8), 2) from differentiated HL-60 cells in a resting condition (where PADI4 protein has been induced) (Bars 3-4 and 9-10), and 3) from activated HL-60 cells after A23187 treatment (Bars 5-6 and 11-12). In EXP2, PADI4 was immunoprecipitated from 1) undifferentiated HL60s (Bars 1-2), 2) differentiated HL60s (Bars 3-4), 3) differentiated HL60s primed with LPS (Bars 5-6), and 4) differentiated HL60s primed with LPS and treated with Calcium ionophore (4μM). In EXP3, PADI4 was immunoprecipitated from 1) mES Control-stable cells (Bars 1-2), 2) mES PADI4-stable cells (Bars 3-4), and 3) mES PADI4-stable cells treated with Calcium ionophore (4μM) for 6 hours. (Bars 5-6). Each condition was performed in duplicate. Treatments are also summarized in Figure 6.2.

In both cases, PADI4 was identified as one of the top proteins in both datasets and depleted from the control. For EXP2, a total of 169 proteins, and for EXP3 a total of 255 proteins were co-enriched with a number of known substrates bearing the hallmark of being substrates under RIME

conditions (enriched in the activated conditions) (Figure 6.8). Of these 16 proteins were at least threefold co-enriched across all PADI4-isolated conditions in EXP2, and 23 proteins were at least threefold co-enriched across all PADI4-isolated conditions in EXP3 (Figure 6.8). In general MS/MS experiments generate very long lists of potential candidates, which pose problems for validation. Many fewer proteins were identified using the RIME methodology, which may therefore be an advantage. The significant problem is that no MS/MS experiment is saturating and there is no current way to quantify missing data. The formaldehyde step appears to decrease quality of the MS/MS data obtained overall seen both from lower LFQ intensity values (two orders of magnitude lower) and given that fewer peptides contribute to protein groups across the analysis (Figure 6.8). Another better approach for transient interaction identification, which was published towards the end of my PhD might be the improved BioID method[10] called TurboID[11] and might retain the MS/MS data quality. The latest paper showed much faster enzymatic modification of the conjugated biotin-ligase enabling proximity ligation within 10 minutes. This would suit the fast timescales of the cell treatment required in this instance.

6.5.3 Differential interactome analysis of EXP5 identifying allosteric PADI4 interactors

MS/MS was then conducted on the new PADI4 activating stimuli identified in Chapter 4 to look for protein interactions that may activate PADI4. I selected the activation condition that made use of GSK3 inhibition in mouse ES cells for detailed MS/MS analysis (CHIR99021 treatment for 45 minutes in serum) (Figure 4.9). Although this activation condition resulted in lower levels of PADI4 activation than after six hours in KSR2i (Figure 4.8), the advantages appeared to me to be that 1) the cell type and environment are as comparable as possible between treatments, 2) the treatment of 45 minutes is rapid and within the framework of cell signalling, 3) there are no confounding effects of serum withdrawal or changes in growth medium, and

4) the levels of PADI4 are both carefully controlled and independent of the stimulus. CHIR99012 is one of the most specific and potent kinase inhibitors available for any kinase target[12-15]. The simplicity of the system therefore ought to give the best chance of observing real physiological regulators. Successful activation of PADI4 protein was confirmed by Western blotting in parallel samples to those prepared for mass spectrometry (Figure 4.9D).

Pulldown using biotinylated Peptide 7 from Chapter 5 was performed alongside immunoprecipitation with PADI4 antibody. Since the antibody and peptide_7 are unlikely to bind to the same face of PADI4, this should allow as comprehensive a set of interacting proteins as possible to be identified. True regulatory proteins pulled down by peptide 7 may not be visible in the antibody pulldown and vice versa, depending on where they bind.

In order to identify allosteric interactors, I decided to include conditions where PADI4 was pre-incubated with saturating quantities of the irreversible inhibitor Cl-amidine to block the active site. Specific enrichment in this condition ought to enrich for proteins that bind allosterically and this condition therefore provides a way to exclude proteins that bind merely at the active site. Cl-amidine pre-treatment was included for both resting and activated cell treatments. If a candidate drives formation of the activated conformation, blocking the active site ought not to prevent this (given the data with Cl-amidine, and the active site probe specific to the active conformation of PADI4 used in Chapter 5, Figure 5.13).

Four conditions were included using three technical replicates each resulting in four treatment groups using PADI4-stable mES cells cultured 1) in serum, 2) in serum + 3μM CHIR99021, 3) in serum + 200μM Cl-amidine, and 4) in serum + 200μM Cl-amidine + 3μM CHIR99021. The experiment was performed in full once with the antibody, and once with the biotinylated peptide 7 to pulldown human PADI4. Dr Christophorou assisted with some sample processing after lysis for handling the 36 replicates for these AP-

MS/MS pulldowns. The data for the antibody and peptide 7 were analyzed separately. Significant differentially interacting proteins across the four conditions were tabulated according to their rank separately for antibody and for peptide_7 (Figure 6.10 and 6.13).

6.5.4 Differential interactome analysis of PADI4 in EXP5

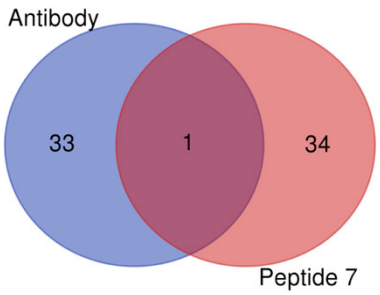
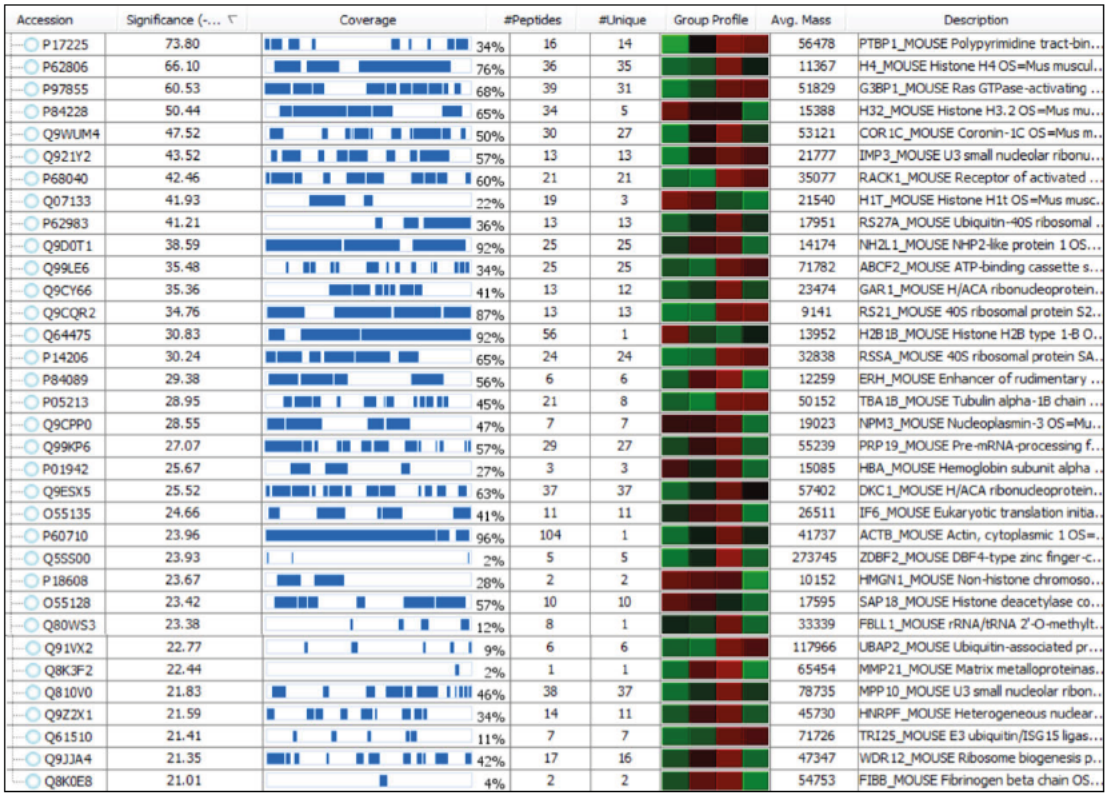


Figure 6.9: Venn diagram of proteins identified between the two pulldowns and with proteins from RIME mES control-stable cells.

The data for the differential interactome of PADI4 are presented separately for the antibody (34 candidates) and for peptide 7 (35 candidates) in order of significance ($10\log_{10}(p) > 20$ corresponds to a p-value < 0.01, and > 40 to a p-value < 0.0001) (Figure 6.10 and 6.13). Significance corresponds both to greatest differences across cell treatments and high confidence detection of its peptide features (Section 6.2.3). One protein (Actin B) was in common between both pulldowns, giving a list of 88 significant candidate regulatory proteins and 29 candidates with a significance of >40 (Figure 6.9 and 6.10). The data for the antibody and the Peptide 7 are presented with the coverage of each protein, the number of peptide features contributing to its detection, and the enrichment or depletion across the different cellular treatments (Figure 6.10 and 6.13).



Accession	Significance (-...	Coverage	#Peptides	#Unique	Group Profile	Avg. Mass	Description
P17225	73.80	34%	16	14		56478	PTBP1_MOUSE Polypyrimidine tract-bin...
P62806	66.10	76%	36	35		11367	H4_MOUSE Histone H4 OS=Mus muscul...
P97855	60.53	68%	39	31		51829	G3BP1_MOUSE Ras GTPase-activating ...
P84228	50.44	65%	34	5		15388	H32_MOUSE Histone H3.2 OS=Mus mu...
Q9WUM4	47.52	50%	30	27		53121	COR1C_MOUSE Coronin-1C OS=Mus m...
Q921Y2	43.52	57%	13	13		21777	IMP3_MOUSE U3 small nucleolar ribonu...
P68040	42.46	60%	21	21		35077	RACK1_MOUSE Receptor of activated ...
Q07133	41.93	22%	19	3		21540	H1T_MOUSE Histone H1t OS=Mus musc...
P62983	41.21	36%	13	13		17951	RS27A_MOUSE Ubiquitin-40S ribosomal ...
Q9D0T1	38.59	92%	25	25		14174	NH2L1_MOUSE NHP2-like protein 1 OS...
Q99LE6	35.48	34%	25	25		71782	ABCF2_MOUSE ATP-binding cassette s...
Q9CY66	35.36	41%	13	12		23474	GAR1_MOUSE H/ACA ribonucleoprotein...
Q9CQR2	34.76	87%	13	13		9141	RS21_MOUSE 40S ribosomal protein S2...
Q64475	30.83	92%	56	1		13952	H2B1B_MOUSE Histone H2B type 1-B O...
P14206	30.24	65%	24	24		32838	RSSA_MOUSE 40S ribosomal protein SA...
P84089	29.38	56%	6	6		12259	ERH_MOUSE Enhancer of rudimentary ...
P05213	28.95	45%	21	8		50152	TBA1B_MOUSE Tubulin alpha-1B chain ...
Q9CPP0	28.55	47%	7	7		19023	NPM3_MOUSE Nucleoplasmin-3 OS=Mu...
Q99KP6	27.07	57%	29	27		55239	PRP19_MOUSE Pre-mRNA-processing f...
P01942	25.67	27%	3	3		15085	HBA_MOUSE Hemoglobin subunit alpha ...
Q9ESX5	25.52	63%	37	37		57402	DKC1_MOUSE H/ACA ribonucleoprotein...
O55135	24.66	41%	11	11		26511	IF6_MOUSE Eukaryotic translation initia...
P60710	23.96	96%	104	1		41737	ACTB_MOUSE Actin, cytoplasmic 1 OS=...
Q5SS00	23.93	2%	5	5		273745	ZDBF2_MOUSE DBF4-type zinc finger-c...
P18608	23.67	28%	2	2		10152	HMGN1_MOUSE Non-histone chromoso...
O55128	23.42	57%	10	10		17595	SAP18_MOUSE Histone deacetylase co...
Q80WS3	23.38	12%	8	1		33339	FBLL1_MOUSE rRNA/tRNA 2'-O-methylt...
Q91VX2	22.77	9%	6	6		117966	UBAP2_MOUSE Ubiquitin-associated pr...
Q8K3F2	22.44	2%	1	1		65454	MMP21_MOUSE Matrix metalloproteinas...
Q810V0	21.83	46%	38	37		78735	MPP10_MOUSE U3 small nucleolar ribon...
Q9Z2X1	21.59	34%	14	11		45730	HNRPF_MOUSE Heterogeneous nuclear...
Q61510	21.41	11%	7	7		71726	TRI25_MOUSE E3 ubiquitin/ISG15 ligas...
Q9JJA4	21.35	42%	17	16		47347	WDR12_MOUSE Ribosome biogenesis p...
Q8K0E8	21.01	4%	2	2		54753	FIBB_MOUSE Fibrinogen beta chain OS...

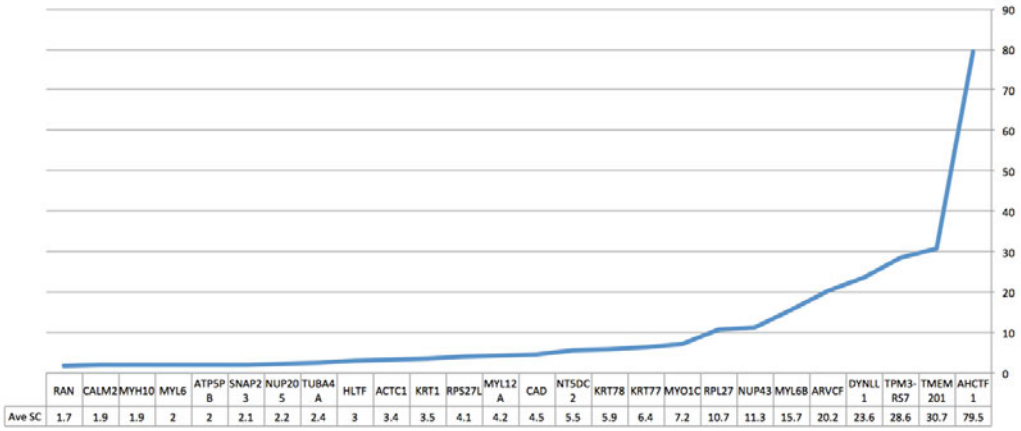
Accession	Significance (-...	Coverage	#Peptides	#Unique	Group Profile	Avg. Mass	Description
Q61879	108.55	68%	196	142		228994	MYH10_MOUSE Myosin-10 OS=Mus musc...
Q60605	104.01	78%	13	11		16930	MYL6_MOUSE Myosin light polypeptide 6 ...
P0DP28	94.75	74%	13	13		16838	CALM3_MOUSE Calmodulin-3 OS=Mus mu...
Q6ZWQ9	87.96	56%	12	12		19895	Q6ZWQ9_MOUSE MCG5400 OS=Mus mus...
Q9WTI7	81.82	54%	59	54		121944	MYO1C_MOUSE Unconventional myosin-I...
P04104	71.34	9%	9	3		65606	K2C1_MOUSE Keratin, type II cytoskeleta...
E9Q0F0	66.87	2%	5	2		112265	E9Q0F0_MOUSE Keratin 78 OS=Mus mus...
P60710	66.25	83%	47	1		41737	ACTB_MOUSE Actin, cytoplasmic 1 OS=M...
Q3UV17	65.43	10%	11	1		62845	K22O_MOUSE Keratin, type II cytoskelet...
A0A0J9YUD5	61.29	15%	31	29		233096	A0A0J9YUD5_MOUSE Nucleoporin 205 OS...
P98203	54.89	2%	2	2		105066	ARVC_MOUSE Armadillo repeat protein de...
P61358	53.74	70%	17	17		15798	RL27_MOUSE 60S ribosomal protein L27 ...
Q6IFZ6	48.77	12%	13	3		61359	K2C1B_MOUSE Keratin, type II cytoskelet...
P62827	45.11	27%	6	6		24423	RAN_MOUSE GTP-binding nuclear protein ...
D3Z2H9	44.34	89%	35	15		28992	D3Z2H9_MOUSE Tropomyosin 3, related s...
Q922U2	43.64	36%	34	17		61767	K2C5_MOUSE Keratin, type II cytoskeleta...
P59235	43.01	12%	4	4		41990	NUP43_MOUSE Nucleoporin Nup43 OS=M...
Q8CI43	42.82	23%	5	3		22749	MYL6B_MOUSE Myosin light chain 6B OS=...
P68033	40.57	40%	31	3		42019	ACTC_MOUSE Actin, alpha cardiac muscle...
B2RQC6	40.56	22%	43	39		243236	PYR1_MOUSE CAD protein OS=Mus musc...
O08638	39.57	15%	42	2		227026	MYH11_MOUSE Myosin-11 OS=Mus musc...
Q6PCN7	35.49	4%	3	3		113317	HLTF_MOUSE Helicase-like transcription f...
P63168	35.26	52%	4	4		10366	DYL1_MOUSE Dynein light chain 1, cytopl...
A2A8U2	34.03	3%	2	2		72500	TM201_MOUSE Transmembrane protein 2...
E9PVG8	33.41	0%	1	1		280230	E9PVG8_MOUSE RIKEN cDNA 9530053A0...
Q6ZWY3	31.84	31%	2	1		9477	RS27L_MOUSE 40S ribosomal protein S27...
G3X9L6	29.26	76%	13	12		18621	G3X9L6_MOUSE ATP synthase subunit d, ...
P68368	28.28	60%	26	4		49924	TBA4A_MOUSE Tubulin alpha-4A chain OS...
Q91X76	28.09	18%	7	7		46034	Q91X76_MOUSE 5'-nucleotidase domain-c...
Q9CQQ7	28.07	37%	10	10		28949	AT5F1_MOUSE ATP synthase F(0) comple...
Q9R0M6	27.91	21%	4	3		22910	RAB9A_MOUSE Ras-related protein Rab-...
Q8CJF7	26.49	10%	21	20		247644	ELYS_MOUSE Protein ELYS OS=Mus musc...
E9Q3T0	25.89	29%	2	2		11433	E9Q3T0_MOUSE Predicted pseudogene 1...
Q61001	24.39	0%	2	2		404056	LAMA5_MOUSE Laminin subunit alpha-5 O...
O09044	24.32	21%	3	3		23261	SNP23_MOUSE Synaptosomal-associated ...

Figure 6.10: List of proteins identified from EXP5. Columns comprise Uniprot accession ID, significance ($-10\log_{10}p$), coverage, number of peptide features, number of unique peptides, enrichment and depletion profiles, average MW and gene description. The order of

treatments in the profile of enrichment and depletion is Group 1: vehicle treatment for 45 minutes. Group 2: activation with GSK3 inhibitor treatment for 45 minutes. Group 3: inhibited with Cl-amidine then vehicle treatment for 45 minutes. Group 4: inhibited with Cl-amidine and then activated with GSK3 inhibitor treatment for 45 minutes. Data used 3 replicates per condition. Full treatments are in Figure 6.2. Green shows depletion, red shows enrichment as for Figure 6.5 as a representation of the $\log_2$(ratio) from -4 to +4. **Top panel:** Antibody pull down. **Bottom panel:** Peptide 7 pulldown.

6.5.5 Background set analysis of EXP5: the CRAP-ome

To validate the candidates, I then sought to analyze the likelihood these proteins might derive from artefactual non-specific binding. A standard use of MS/MS negative controls might be to use beads only or an isotype control to attempt to capture a background set of non-specific interactors. The logic of this approach is somewhat flawed - the true background set that we are most interested in obtaining is anything that is not regulatory (doesn't differ between cellular treatments) rather than anything that can be enriched by non-specific binding. A sticky protein might bind beads alone more intensely than PADI4 isolated on beads, but still be differentially bound to the PADI4 in resting versus activated cells. Clearly it should not be excluded as a non-specific interaction merely because it interacts strongly with beads. Considering background binding is of course nonetheless important, so to avoid these problems, I considered that tempering possibly significant differential regulatory candidates with reference to the Contaminant Repository for Affinity Purification–Mass Spectrometry (CRAP-ome) would be a more useful method, alongside retrospective comparison with a separate negative set obtained in these cell lines[16] (Figure 6.11 and 6.12). This approach avoids excluding data *de facto* from inappropriate prediction of the MS/MS background set.



	RAN	CALM2	MYH10	MYL6	ATP5PB	SNAP23	NUP205	TUBA4A	HLTF	ACTC1	KRT1	RPS27L	MYL12A	CAD	NT5DC2	KRT78	KRT77	MYO1C	RPL27	NUP43	MYL6B	ARVCF	DYNLL1	TPM3-RS7	TMEM201	AHCTF1
Ave SC	1.7	1.9	1.9	2	2	2.1	2.2	2.4	3	3.4	3.5	4.1	4.2	4.5	5.5	5.9	6.4	7.2	10.7	11.3	15.7	20.2	23.6	28.6	30.7	79.5

Figure 6.11: List of proteins identified from the Peptide 7 pulldown, with homologues in Homo sapiens, were queried against the CRAP-ome database. The average spectral counts across all 400 background experiments is plotted.

To do this, I took a list of the candidate mouse proteins and generated a list of their human homologues that I queried against the CRAP-ome database (this database does not have data for *Mus musculus* yet). I then plotted the average spectral counts for each of the candidates across the 400 background experiments contained within the CRAP-ome dataset (Figure 6.11). For several promising candidates, I have in addition presented the distribution of spectral counts across the background set experiments (Figure 6.12) as well as presenting two candidates that you might ordinarily expect to be from the background set for comparison (Krt1 and Actin B). Given that PADI4 is known to take keratins and actins as substrates, the classic background set profiles of spectral counts for Krt1 and Actin B may or may not indicate that they are true false positives. Some of the most promising high confidence candidates do not have this profile, however, which is a fair indication they derive from specific binding to PADI4 (Figures 6.11 and 6.12).

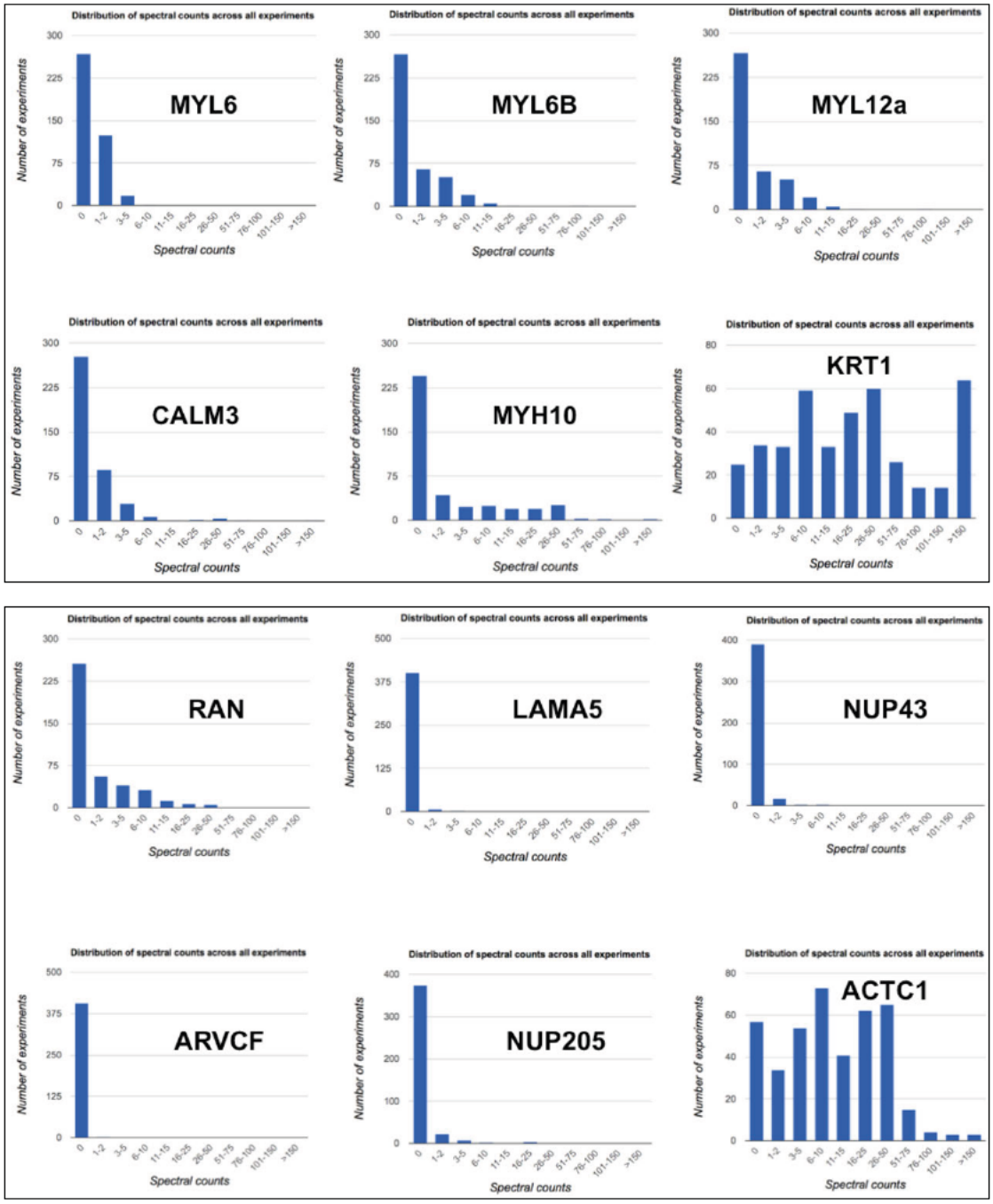
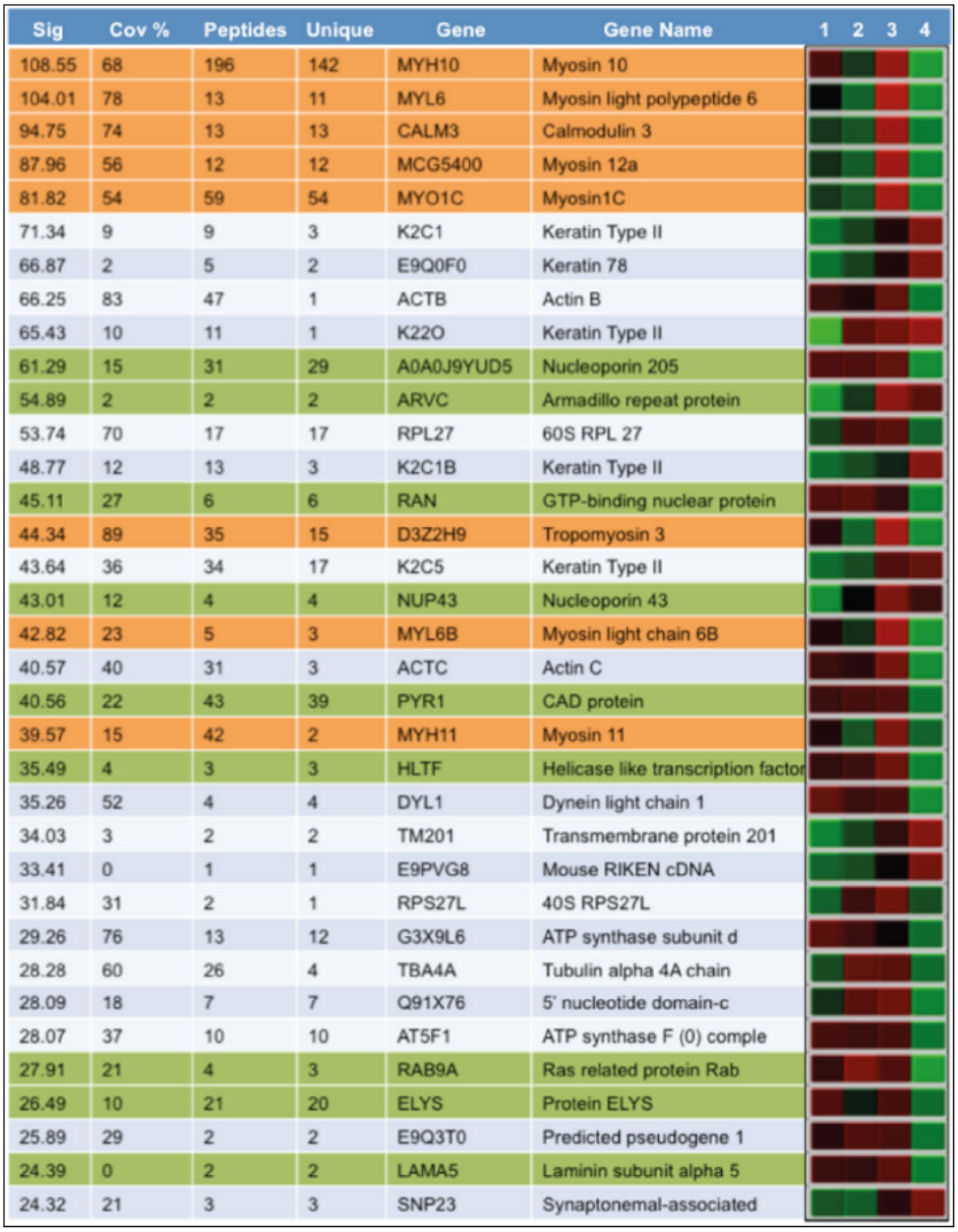


Figure 6.12: Background binding profiles of a selection of promising candidates identified from Peptide 7 pulldown. Background binding profiles of **h**omologues from *Homo sapiens* are shown with the distribution of their spectral counts identified across the 400 background set experiments. If a protein appears in a large number of background experiments, it is more likely to appear in negative control AP-MS/MS experiments (such as the profiles for ACTC1 and KRT1). Conversely, LAMA5 has a spectral count of zero in >375 experiments. **Top panel:** Protein candidates with known calcium binding functions (orange, Figure 6.13). **Bottom panel:** Nuclear protein candidates (green, Figure 6.13).

6.5.6 Candidate interacting proteins for the regulation of PADI4

The dataset provides some very interesting protein candidates that bind to a different extent between cellular treatments. From the peptide 7 data in particular several of the top candidates leapt out as being potential regulators for their previously known calcium binding functions. Eight proteins showed a similar pattern of differential enrichment and these are highlighted in orange in Figure 6.13 (detail of Figure 6.9, bottom panel). Three of these were small proteins with multiple EF-hand domains (a prototypical calcum-binding domain) including calmodulin, MYL6 and MYL6B. In addition to these EF-hand containing proteins, a number of calmodulin-binding proteins were identified with a similar profile of enrichment and depletion across the different cellular treatments (each of which possess at least one calmodulin-binding IQ-motif: Myosin 10, Myosin 12a, Myosin 1C, Tropomyosin 3 and Myosin 11). Most interesting was that this group of proteins was significantly enriched in conditions where PADI4 was at a resting state and where the PADI4 active site had been blocked by preincubation with Cl-amidine (Figure 6.13, Group 3). In the activated conditions, these proteins were found to be depleted (Figure 6.13, Group 4).

These data evoke a possible hypothesis where EF-hand containing proteins interact with PADI4 in the resting condition. As these proteins appear to be enriched in conditions where the PADI4 active site is blocked in the resting state, they are therefore likely to be interacting allosterically, which is consistent with other activating molecules such as the peptides in Chapter 5 and the cross reactive antibodies from *Darrah et al.*[17]. What is less clear is the explanation of the reduced intensity of calcium binding protein interactors to activated PADI4 – one possibility is that an activating stimulus could disrupt the interaction between these calcium-binding proteins and PADI4, thereby elevate the local calcium concentration, and release activate PADI4.



Sig	Cov %	Peptides	Unique	Gene	Gene Name	1	2	3	4
108.55	68	196	142	MYH10	Myosin 10				
104.01	78	13	11	MYL6	Myosin light polypeptide 6				
94.75	74	13	13	CALM3	Calmodulin 3				
87.96	56	12	12	MCG5400	Myosin 12a				
81.82	54	59	54	MYO1C	Myosin1C				
71.34	9	9	3	K2C1	Keratin Type II				
66.87	2	5	2	E9Q0F0	Keratin 78				
66.25	83	47	1	ACTB	Actin B				
65.43	10	11	1	K22O	Keratin Type II				
61.29	15	31	29	A0A0J9YUD5	Nucleoporin 205				
54.89	2	2	2	ARVC	Armadillo repeat protein				
53.74	70	17	17	RPL27	60S RPL 27				
48.77	12	13	3	K2C1B	Keratin Type II				
45.11	27	6	6	RAN	GTP-binding nuclear protein				
44.34	89	35	15	D3Z2H9	Tropomyosin 3				
43.64	36	34	17	K2C5	Keratin Type II				
43.01	12	4	4	NUP43	Nucleoporin 43				
42.82	23	5	3	MYL6B	Myosin light chain 6B				
40.57	40	31	3	ACTC	Actin C				
40.56	22	43	39	PYR1	CAD protein				
39.57	15	42	2	MYH11	Myosin 11				
35.49	4	3	3	HLTF	Helicase like transcription factor				
35.26	52	4	4	DYL1	Dynein light chain 1				
34.03	3	2	2	TM201	Transmembrane protein 201				
33.41	0	1	1	E9PVG8	Mouse RIKEN cDNA				
31.84	31	2	1	RPS27L	40S RPS27L				
29.26	76	13	12	G3X9L6	ATP synthase subunit d				
28.28	60	26	4	TBA4A	Tubulin alpha 4A chain				
28.09	18	7	7	Q91X76	5' nucleotide domain-c				
28.07	37	10	10	AT5F1	ATP synthase F (0) comple				
27.91	21	4	3	RAB9A	Ras related protein Rab				
26.49	10	21	20	ELYS	Protein ELYS				
25.89	29	2	2	E9Q3T0	Predicted pseudogene 1				
24.39	0	2	2	LAMA5	Laminin subunit alpha 5				
24.32	21	3	3	SNP23	Synaptonemal-associated				

Figure 6.13: Mass spectrometry data from PADI4-stable cells are displayed with the top differentially detected proteins shown with analysis performed using Peaks7.0 software. For each condition, three replicate experiments were performed. PADI4 was isolated using peptide_7 from PADI4 stables in a resting state, after treatment with CHIR99021 for 45 mins, after Cl-amidine treatment for 30 minutes in a resting condition, and after Cl-amidine treatment for 30 mins combined with treatment with CHIR99021 for 45 mins. Significance ($-10\log_{10}(P)$) is given in column 2 where a significance of 20 corresponds to a p-value of 0.01 (Section 6.2.3). Protein coverage is displayed in column 3. The number of peptides and number of unique peptides detected for each protein is given in column 4 and 5. Log_2 enrichment of the protein for each grouped condition (combined across the three replicates) is displayed as a heatmap in column 6. Eight proteins showing the same profiles

of enrichment (specifically enriched in the Cl-amidine treated resting condition) are highlighted in orange. These proteins are particularly interesting as they are either calcium or calmodulin binding proteins. A further nine nuclear-located proteins are highlighted in green.

A second set of candidates were highlighted (Figure 6.13 in green) as they are known to be located in the nucleus (such as on UniProtKB) or specifically involved in nuclear transport (RAN, NUP205, NUP43, Protein ELYS, LAMA5). This is interesting with respect to PADI4 as it has a nuclear localization signal and has been well characterized for functions in the nucleus (such as in histone modification)[4,18-20]. A recently published paper analyzing PADI2 using proximity ligation with BioID2 identified RAN as a PADI2 interactor[6]. This study showed RAN binding was involved in calcium dependent nuclear localization. These nuclear candidates therefore offer further avenues for future work.

Another notable candidate appeared from the antibody dataset, which was RACK1 (Receptor of activated protein C kinase 1) (Figure 6.10, top panel). This is particularly promising in light of a previous paper which showed regulation of PADI4 after use of small molecule PKC inhibitors[21]. From my MS/MS data, RACK1 interacts allosterically (in the Cl-amidine treated conditions), but not differentially between activated and resting treatments. It is conceivable that RACK1 could act as a scaffold for regulation of PADI4 by activated PKC isoforms and this would be interesting to explore further.

6.6 Validating calmodulin as a candidate regulator of PADI4

6.6.1 Choosing a candidate activator from MS/MS data

From literature searches at the beginning of the project, the calcium regulatory protein calmodulin (CaM) had stood out as a possible activating candidate for calcium regulated activation. Calmodulin responds to physiological calcium increases by changing its conformation dramatically on calcium binding (Figure 6.14A and B). Interestingly, it has been shown previously to regulate kinase and phosphatase catalytic activity by calcium-dependent binding. It is also involved in regulating many other cellular

proteins and is a key transducer of calcium-dependent cellular signalling. After it was identified as one of the most significant novel interacting candidates from EXP5, it was the first protein that I decided to validate from the MS/MS dataset.

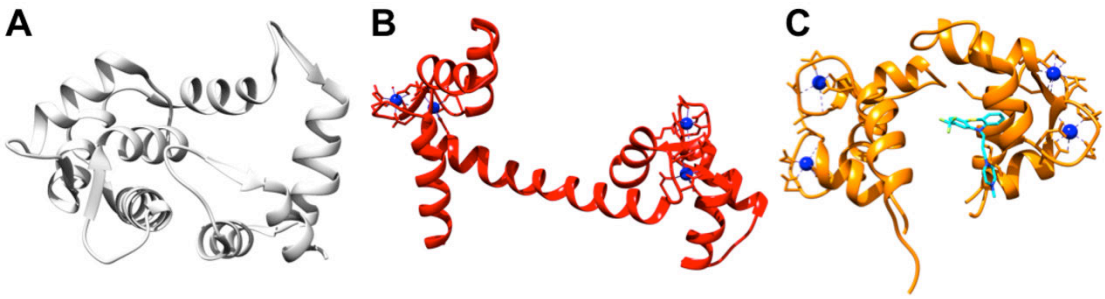


Figure 6.14: Crystal structures of calmodulin in **A:** the apocalmodulin structure, shown coloured in white, in the absence of calcium binding (pdb: 1qx5) **B:** the calmodulin structure in the presence of full calcium occupancy shown coloured in red (pdb: 3cln) **C:** the inhibited calmodulin structure, shown coloured in orange, with full calcium occupancy and bound to the small molecule TFP (coloured in cyan) (pdb: 1ctr). Calcium ions are represented as dark blue spheres.

In EXP5, calmodulin was identified as the third most significant differentially interacting protein with PADI4, and was enriched in conditions where the active site had been blocked with Cl-amidine, suggesting the interaction may be allosteric. I then retrospectively looked more closely at the other experiments to see if calmodulin could be identified in the full MaxQuant data analysis. EXP1 also contained calmodulin in the full list of interacting proteins where it was identified as enriched in the resting condition but depleted from the activated condition. This adds further some support to the MS/MS data from EXP5 as calmodulin could be identified in both HL-60s and in mES cells and therefore across two different cell types, two different species and two different PADI4 activating stimuli.

6.6.2 PADI4 binds calmodulin by reciprocal pull-down

It was therefore attempted to confirm the interaction identified by MS/MS by performing the reciprocal pull-down of calmodulin and assessing if PADI4 could be co-enriched. This work was done with Emma Clarke (a PhD rotation

student in the lab) and Dr Christophorou performed the repeat of the experiment (Figure 6.15). PADI4-stable mES cells, alongside PADI4-null mES cells, were transfected with CaM-GFP. GFP-trap beads were then used to isolate the exogenous calmodulin from the lysate, eluted and run for Western blot with detection by anti-human PADI4 antibody. Given the MS/MS data indicated a difference between resting and active conditions, the pulldown was performed firstly in the presence of Ca^{2+} and secondly in the presence of the calcium chelator EDTA. Results showed that human PADI4 is enriched by pull-down of CaM-GFP with more efficient co-enrichment in the absence of Ca^{2+} (Figure 6.15). These data are consistent with the results from the mass spectrometry data and suggests the biochemical interaction of calmodulin with PADI4 can be recapitulated from reciprocal pulldown of calmodulin.

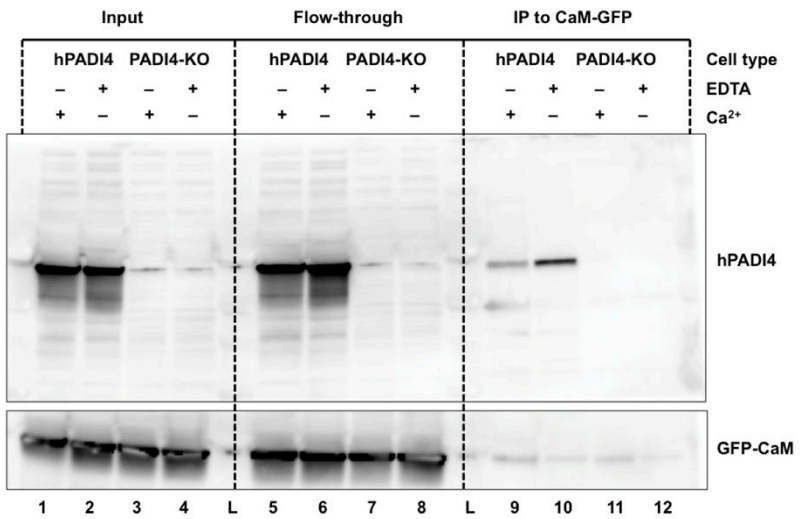


Figure 6.15: Reciprocal pulldown with calmodulin confirms PADI4 interaction. Immunoblot analysis of human PADI4 and GFP after CaM-GFP trap pull-down performed in the presence of 5mM EDTA (Lanes 2, 4, 6, 8, 10, 12) or 5 mM $CaCl_2$ (Lanes 1, 3, 5, 7, 9, 11). PADI4-stable mES cells (Lanes 1-2, 5-6, 9-10) or PADI4-null mES cells (Lanes 3-4, 7-8, 11-12) were transfected with GFP-CaM before pull-down. This experiment was undertaken with the help of Emma Clarke; this figure is from a repeat of the experiment, which was performed by Dr Christophorou. Data are representative of n = 2.

A couple of future experiments will be very interesting to develop these data. Firstly, a further pull-down using GFP-beads from cells that have not been

transfected with GFP-CaM will be a useful additional control. Secondly, performing the reciprocal pull-down in the presence of Cl-amidine would be useful to confirm that the interaction with calmodulin is not mediated through the active site of the enzyme. Additionally, performing this pull-down after treatment with activating peptides 11 and 14 would show whether calmodulin and the peptides bind at the same location in the enzyme, if for example, the peptides were to prevent the interaction with calmodulin.

6.6.3 Recombinant CaM activates PADI4 *in vitro*

Given these data and the previous roles for calmodulin in calcium dependent enzyme regulation, I then performed a citrullination lysate assay to test if recombinant bovine calmodulin might increase PADI4 activity at limiting calcium concentrations, analogously to the activating peptides (Chapter 5). To do this, 98 µL clarified PADI4-stable cell lysate was added to different assay tubes. Recombinant calmodulin (rCaM) or vehicle was added to the lid of treated tubes. A serially diluted Ca^{2+} concentration was added to the lid of each tube to cover the range of limiting calcium dependence shown by PADI4 *in vitro* (up to 100 µL total assay volume). The tubes were spun quickly to start the assay and incubated at 37˚C with detection by Western blot to citrullinated H3 for exactly 30 mins. This enabled a tightly controlled comparison of activation with and without the addition of rCaM.

Addition of rCaM did not elicit a PADI4 activation response (Figure 6.15A). In the presence of Ca^{2+}, therefore, rCaM does not appear to activate PADI4 *in vitro*. In addition, these data suggest that rCaM, activated by Ca^{2+}, does not activate a component in the lysate that acts on PADI4 either. Then, given that calmodulin bound more strongly in the inactive condition (by MS/MS and by reciprocal pull-down), I tested the possibility that rCaM may only activate PADI4 if it can first interact with the inactive conformation of PADI4. I therefore repeated the assay, but preincubated rCaM in the lysate with PADI4 in the presence of EDTA to quench any calcium already present in the cell lysate. After 20 mins of preincubation, $CaCl_2$ was then added to the tube

lids as before to a concentration that accounts for the added EDTA (EDTA chelates calcium in a 1:1 ratio) but which then surpasses this level to give the same limiting concentrations of Ca^{2+} as in the previous assay (500μM + 100/250μM $CaCl_2$). In principle, this would allow calmodulin to bind to the inactive conformation of PADI4 before the citrullination assay was initiated with calcium. Excitingly, PADI4 was now activated to a greater extent in the conditions where rCaM had been added (Figure 6.15B-C). The assay was repeated in the presence of trifluoperazine (TFP) during the pre-incubation step, an irreversible small molecule inhibitor of CaM (crystal structure of inhibition is shown in Figure 6.13C). TFP attenuated the CaM-mediated PADI4 activation, without inhibiting PADI4 activation in the other conditions. This suggests a specific interaction of apocalmodulin with PADI4, which acts to activate PADI4 at a reduced Ca^{2+} concentration and is consistent with the MS/MS data.

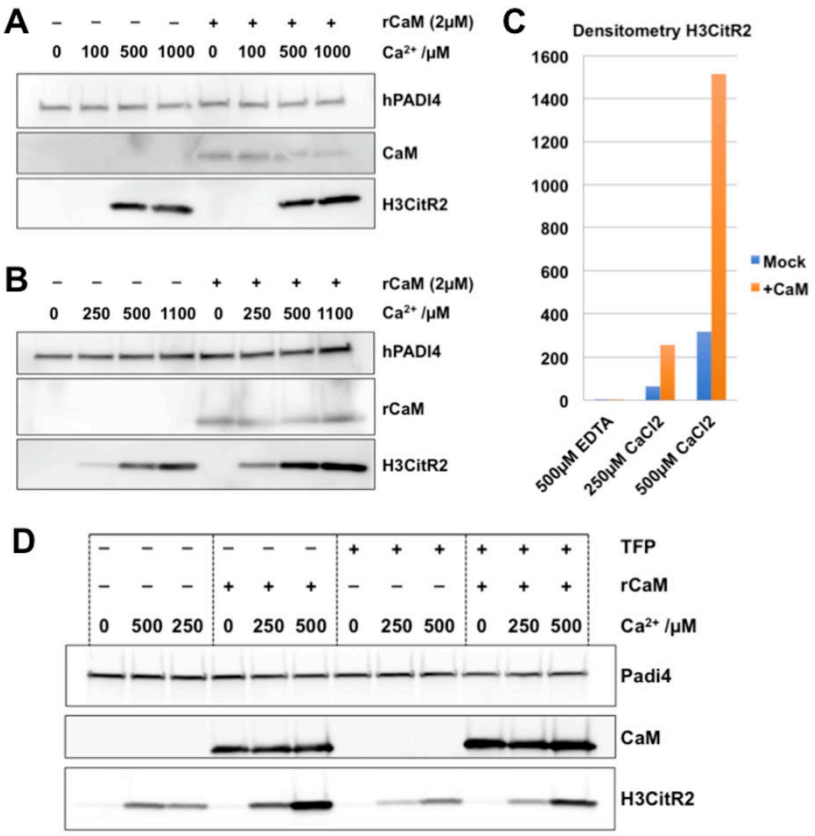


Figure 6.16: Recombinant Calmodulin activates PADI4 *in vitro*. A: Immunoblot analysis of H3CitR2 and CaM of a lysate citrullination assay using PADI4-stable mES cells. Human PADI4 is presented as a loading control. **A:** Lysates were supplemented with vehicle (Lanes

1-4) or recombinant bovine Calmodulin (Lanes 5-8) in the presence of serially diluted $CaCl_2$ and incubated for 30 mins at 37˚C before blotting. **B:** Lysates were pre-incubated with vehicle (Lanes 1-4) or recombinant bovine Calmodulin (Lanes 5-8) in the presence of EDTA and incubated for 20 mins at 37˚C. Lysates were then supplemented with serially diluted $CaCl_2$ and incubated for a further 30 mins at 37˚C before blotting. **C:** Densitometry analysis of conditions shown in panel B normalized to PADI4 levels performed in ImageJ. **D:** Immunoblot analysis of H3CitR2 and CaM of a lysate citrullination assay using PADI4-stable mES cells. Human PADI4 is presented as a loading control. Lysates were supplemented with vehicle (Lanes 1-3, 7-9) or recombinant bovine Calmodulin (Lanes 4-6, 10-12) in the presence of EDTA, treated with vehicle (Lanes 1-6) or TFP (Lanes 7-12), and incubated for 20 mins at 37˚C. Lysates were then supplemented with serially diluted $CaCl_2$ and incubated for a further 30 mins at 37˚C before blotting. For Panels A and B, data are representative of n = 2; for the additional controls in Panel D, data are shown from a single preliminary experiment.

Similar experiments to those performed in Chapter 5 (Figure 5.13) using biotinylated F-amidine, a reagent that reacts only with the activated PADI4 conformation would be good as a follow-up. Incubating recombinant calmodulin and recombinant PADI4 in the presence of an increasing concentration of calcium, before treatment with biotinylated F-amidine would reveal whether the presence of calmodulin could drive formation of an activated conformation of PADI4 at lower calcium concentrations.

6.6.4 Putative CaM binding site on PADI4

Given the broad roles of calmodulin in regulating calcium signalling in the cell, I therefore considered that disruption of calmodulin and PADI4 through targeted perturbations made to PADI4 would be most amenable in deducing a cellular role for calmodulin activation of PADI4. To explore this line of enquiry more closely, bioinformatic approaches were used to identify possible calmodulin binding sites on PADI4 (using the Calmodulin target database, CaMELs and Calmodulation tools)[22-25]. Multiple putative calmodulin binding site motifs on PADI4 were identified with one particularly good candidate that arose from multiple computational methods (with a 1-8-14 motif) and was consistent in sequence composition with CaM binding sites that have previously been identified in proteins that bind the calcium-

unbound conformation of CaM (in charge and secondary structure: with +3 overall net charge)[26]. These regions were compared against sequences in other species to assess conservation (Figure 6.16). The region is conserved in PADI4 paralogues. Interestingly the region is also conserved in mammalian PADI2, but is deleted in cyanobacterial protein. In addition, point mutants of this putative CaM binding region in PADI2 allow full calcium activation in the crystal structure (F220A and F221A)[27]. This is suggested to allow full calcium occupancy in the crystal structure by disrupting the dimer interface. A recent paper analyzing interactions of PADI2 found this region bound to RAN and modulated nuclear transport[6]. It is notable that RAN was also identified as one of my most significant differential interacting proteins to PADI4 (Figure 6.12). Lastly, a paper identified that auto-antibodies to PADI4 are found to specifically bind at this region[28], not far from where activating antibodies were hypothesized to bind[17].

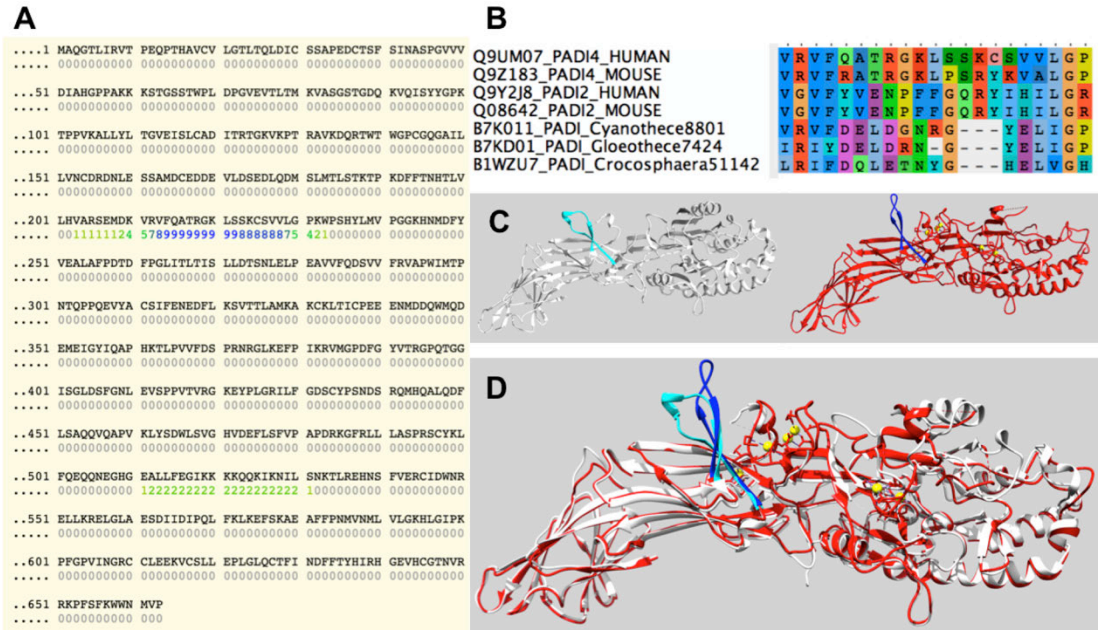


Figure 6.17: Putative Calmodulin binding motif on PADI4. **A:** The top candidate region from the calmodulin target database is also identified from the calmodulation as the following overlapping target motifs: 1-8-14, FQATRGKLSSKCSV, 214-227, 1-16, FQATRGKLSSKCSVVL, 214-229 and 1-14, FQATRGKLSSKCSV, 214-227. **B:** PADIs from three cyanobacteria and from human and mouse PADI4 and PADI2 were aligned using Mafft L-ins-I and the putative calmodulin binding motif region was extracted and viewed in AliView. Cyanobacterial proteins show a truncation in this region. **C** and **D:** The region is unstructured in the PADI4 crystal structure (pdb: 1wd8 and 1wd9), but is conserved in PADI2 where it is

ordered (pdb: 4n2c) and changes conformation between inactive and active conformations. Inactive and active conformations (white, 4n20 and red, 4n2c) of PADI2 were superimposed using Chimera using MatchMaker (Section 2.5.3.4). The region highlighted in cyan (4n20) or blue (4n2c) is the location of the putative CaM motif.

This motif is located adjacent to the Ca3-5 binding switch and is an unstructured region in the protein crystal structure in PADI4. In the holoenzyme of PADI2 is the fully active conformation generated, in which this CaM binding region is shown to be ordered (Figure 6.16B)[27]. This is shown in Figure 6.18. In PADI4 the region remains disordered in both structures so it is currently unknown how this region affects activation and may explain how this has evaded previous understanding elicited from published crystal structures.

As other good calmodulin binding motifs also exist on PADI4, it will be important to obtain more data as to calmodulin regulation of other paralogues in vitro. In particular, it is important to test whether recombinant calmodulin activates PADI4, PADI2, and cyanoPADI proteins. This will help determine whether the calmodulin binding region is expected to be evolutionarily conserved across the three isozymes and narrow down which CaM binding motifs are most promising for the design of point mutants (Discussed in the section below).

6.6.5 Discussion of immediate follow-up work on calmodulin

The data showing calmodulin activation in vitro point to immediate follow-up experiments. independently from cell lysates. One important test will be whether recombinant calmodulin activates recombinant PADI4, PADI2, and cyanoPADI proteins outside of a cell lysate context. This would provide further support for a role of calmodulin in activating PADI4 by direct binding (as implied by the MS/MS data and the reciprocal pull-down) rather than for example by acting on an unknown component of the lysate that might subsequently act on PADI4. The data are interesting in either model but it will be important to distinguish between these scenarios.

From this work, point mutants of PADI4 that are unable to bind calmodulin can be readily introduced into cells by site directed mutagenesis of the PiggyBac PADI4-stable vector construct to make calmodulin-binding dead mutants. PADI4 point mutants can then be introduced into mouse embryonic stem cells to generate mutant PADI4 stables (in as little as a week or two) and activity can then be tested both in the lysate assay and using the cellular activation conditions established in Chapter 4. These experiments will provide much stronger evidence in confirming a role for calmodulin in PADI4 activation than perturbing calmodulin globally. This nods to the wider consideration that studying effects of calmodulin in general is not always easy given the many roles CaM plays in regulating calcium fluxes in the cell and the likelihood for toxicity. Mutating the effect at the point of PADI4 protein will allow much stronger and more specific mechanistic conclusions to be drawn with respect to regulation of PADI4 by calmodulin binding. This is especially true given the possibility for cell death to activate PADI4 artefactually such as from the provision of high calcium ion concentrations that would be available in the extracellular space after the cell membrane is compromised.

Given the data showing that the cyanobacterial protein has a more stringent calcium requirement than mammalian PADI4 (Figure 3.11), it will be particularly interesting if rCaM only activates the mammalian isoenzymes and not cyanoPADI. This would be consistent with the calmodulin-binding motif on PADI4 being regulatory for calmodulin dependent binding. While it is fully conserved in PADI4 orthologues and most likely maintained in PADI2 sequences, it is truncated in cyanoPADI. Additionally if calmodulin does not activate cyanoPADI, then the truncation found in the cyanobacterial sequence can be introduced into the mammalian homologue as a model for a calmodulin-dead mammalian PADI mutant by producing the hybrid mutant both recombinantly and in cells.

6.6.6 Calmodulin Discussion

These data, taken together, are consistent with a hypothesis where calmodulin binds to PADI4 in the inactive calcium-unbound state, such that only once calcium is added, calmodulin is released from binding to PADI4. This release event then generates active PADI4, but it is unknown exactly by what mechanism. It is possible this is achieved through elevation of the local calcium concentration either as a result merely of the calmodulin interaction or potentially mediated through additional interacting proteins. From the MS/MS data, it is very interesting that other candidates including a variety of Myosin proteins also bound PADI4 in the inactive resting state and also showed decreased binding in the activated condition. These calcium-binding myosins could also be responsible for activating PADI4 in a similar manner to Calmodulin. As these myosin proteins possess calmodulin-binding motifs (IQ motifs) they may also interact in combination with calmodulin. IQ motifs in particular are known to bind calmodulin in the absence of calcium (apocalmodulin). It is a possible hypothesis, therefore, that if an interaction between calcium-binding myosin proteins is mediated by apocalmodulin, that an increase in calcium could disrupt and change the conformation of apocalmodulin and release calcium bound by myosins to elevate the local concentration.

Also interesting is that many IQ motifs are protein kinase C (PKC) phosphorylation sites; there is extensive calmodulin and PKC crosstalk in the literature[29-35]. This could in principle connect the published roles found for opposing PKC isozymes in regulating PADI4 derived from relatively unspecific small molecule inhibitors[21] and develop the observations of effects from small molecule inhibitors to a clearly defined mechanism at the level of PADI4 protein. Furthermore this could connect to the identification of RACK1 (Receptor of Activated PKC 1) as a PADI4 interacting protein from the MS/MS data (Figure 6.10). These possibilities remain very interesting avenues to explore in future work to elucidate the precise mechanism responsible for PADI activation in cells.

Another interesting consideration is that calmodulin is a well-established myelin basic protein (MBP) interacting protein[36-39]. MBP is the major component of myelin. This points to a connection to the role of PADI2 or PADI4 in MBP citrullination and also to PADI overexpression and hypercitrullination in multiple sclerosis. This may also be interesting with respect to evolutionary conservation especially given the complicated evolutionary origin identified for PADIs in Chapter 3. The modes of binding of myosins and calmodulin have been found to be well-conserved across different species[40 41 42 43]. As PADIs appear to have been acquired horizontally in the animal lineage, co-evolution with specific calcium binding proteins in the vertebrate lineage may have enabled intracellular roles to have evolved. Testing recombinant proteins from various species for their capacity to be activated by recombinant calmodulin is likely to make identifying the CaM binding motif easier, but will also provide exciting avenues towards understanding how physiological PADI regulation may have evolved in the vertebrate lineage.

6.7 Discussion of future MS/MS approaches for PTM identification

6.7.1 Future work to identify S-nitrosylation

Given the indications of possible active site modification, experiments to test the hypothesis that it may be S-nitrosylated would be interesting. Diethylammonium (Z)-1-(N,N- diethylamino)diazen-1-ium-1,2-diolate (DEA NONOate) is a crystalline solid that when dissolved spontaneously dissociates into the free amine and nitric oxide (NO). It therefore acts as a NO donor in solution with a half-life of 2 minutes at 37°C and 16 minutes at 22-25°C to liberate 1.5 moles of NO per mole. NO is highly reactive and will spontaneously nitrosylate cysteine residues *in vitro* that can be modified in an endogenous protein context. Treatment with DEA NONOate may be used to test on recombinant PADI4 to see if citrullination may be disrupted or enhanced. Equally, this approach may be employed on cells to test whether PADI4 activation may be induced or inhibited directly by nitrosylation and

whether the physiological activation conditions identified in Chapter 4 may be disrupted or enhanced by DEA NONOate pre-treatment.

A complementary approach using the 'biotin switch' method could also be used to detect cysteine S-nitrosylation modifications directly either by Western blot or MS/MS[44,45]. Briefly, PADI4 protein is treated with methyl-methane thiosulfonate (MMTS) such that any free cysteines will be methylated and protected, but which leaves nitrosylated cysteines intact. Then, ascorbic acid is used to reduce the S-nitrosylated cysteines, but which is not able to reduce the methylated cysteines. Finally the protein is treated with HPDP-Biotin or an iodo-tandem mass tag (iodoTMT) reagent, that covalently conjugates endogenously S-nitrosylated cysteines (that were subsequently converted to free cysteines) to either biotin, or the TMT group. Detection can then be performed using either Streptavidin-HRP or anti-TMT by Western blot or by using mass spectrometry with the TMT.

The biotin switch method mirrors the speculative inference made in the observation of differential carbamidomethylation efficacy, which had been suggested by the initial mass spectrometry (Figure 6.3). It is hoped these plans, as they could not be undertaken within the timeframe of the PhD, might be undertaken in the near future as all the reagents are ready. That PADI catalysis uses a cysteine in the active site means S-nitrosylation is a particularly interesting possibility. In addition, other members of the pentein-containing fold that PADIs possess (described in detail in Chapter 5) have previously been shown to be regulated by endogenous nitrosylation: in fact DDAH, Argininosuccinate synthetase and ornithine decarboxylase enzymes are all regulated by Cysteine-S-nitrosylation[46-52].

6.7.2 Future work to analyze phosphorylations

Another important avenue to explore in the future is mapping phosphorylations to PADI4. One major difficulty in mapping comprehensive phosphorylations is their stability. Phosphorylations are generally sub-

stoichiometric and labile in a cellular lysate environment where they are subject to enzymatic removal. As an approach to look for phosphorlyations will be particularly interesting for PADI4, it is likely to be profitable to perform phosphopeptide enrichment strategies such as by TiO_2 column enrichment or by using phos-tag reagent enrichment. Discussions with Dr Greg Findlay at the MRC PPU suggested that may be possible through collaboration. A recent approach to identify non-canonical phosphorylation by strong anion exchange chromatography may also be useful[53].

6.8 Concluding Discussion

In this chapter, work was progressed towards understand how PADI4 may be physiologically regulated in cells. Cellular activating conditions from Chapter 4 were subjected to proteomic analysis to produce a list of candidate regulatory events. This included the phosphorylation of Ser433 as part of a putative phosphorylated motif, and a list of interacting proteins that include a number of EF hand containing calcium binding proteins, RACK1 and a number of nuclear proteins. One of the candidates, calmodulin, was validated further by reciprocal pulldown and was shown to activate PADI4 in vitro.

The similarity between S100 proteins and calmodulin is provocative especially with reference to the possible regulation of PADI3/PADI1[54] by S100A3. S100 proteins are thought to have evolved from a calmodulin duplication in the animal lineage and over 21 different proteins exist in humans. This could explain the tissue specific regulation of PADIs. For example, S100A8/A9 proteins are among the most abundant proteins in neutrophils, whereas calmodulin itself is a major interacting protein with MBP at the myelin. It is therefore conceivable that different calcium binding proteins may regulate different PADI paralogues in different cell type specific contexts. It is clear that targeted exploration of EF-hand type calcium binding proteins would be merited from the evidence contained in this chapter.

The final consideration relates to possible connections of calmodulin to other candidate regulators evidence within the Chapter. Firstly calmodulin regulation may relate to S-nitrosylation. Several papers have identified cross talk between calmodulin and S-nitrosylation such as with nitric oxide synthases and enzymes structurally related to the PADIs (containing the same pentein catalytic fold)[46-52]. Secondly there are extensive connections between calmodulin and PKCs that have been documented in the literature. Given the body of evidence suggesting that PADI4 regulation is complicated, explorations of the possibly multi-faceted regulation of PADI4 will be required for a full understanding of their physiological regulation. Requirements for different levels of activation in cells are likely and mechanisms to carefully restrict the drastic consequences caused by hypercitrullination may be different from those required in transcriptional fine-tuning.

As such the work in Chapter 4 and this Chapter 6 provides a platform for the discovery of the precise biological regulation of PADI4 and other PADI paralogues. It remains in particular to identify how the activation signal may be transmitted via calmodulin, given the reduction in binding to PADI4 after activation. This is perhaps most likely to be mediated through phosphorylation given the evidence provided of kinase inhibition regulating PADI4 activation in cells, downstream of canonical Wnt signalling. In addition, it will be particularly interesting to consider further how PADI regulation evolved over evolutionary time and as such to elucidate how the regulation of PADIs evolved in the vertebrate lineage and in different cellular contexts.

6.9 References for Chapter 6

1. Andrade, F. *et al.* Autocitrullination of human peptidyl arginine deiminase type 4 regulates protein citrullination during cell activation. *Arthritis & Rheumatism* **62,** 1630–1640 (2010).
2. Liu, Y.-L., Chiang, Y.-H., Liu, G.-Y. & Hung, H.-C. Functional role of dimerization of human peptidylarginine deiminase 4 (PAD4). *PLoS ONE* **6,** e21314 (2011).
3. Slack, J. L., Causey, C. P., Luo, Y. & Thompson, P. R. Development and Use of Clickable Activity Based Protein Profiling Agents for Protein Arginine Deiminase 4. *ACS Chem. Biol.* **6,** 466–476 (2011).
4. Christophorou, M. A. *et al.* Citrullination regulates pluripotency and histone H1 binding to chromatin. *Nature* **507,** 104–108 (2014).
5. Tilvawala, R. *et al.* The Rheumatoid Arthritis-Associated Citrullinome. *Cell Chem Biol* **25,** 691–+ (2018).
6. Zheng, L. *et al.* Calcium Regulates the Nuclear Localization of Protein Arginine Deiminase 2. *Biochemistry* **58,** 3042–3056 (2019).
7. Giansanti, P., Tsiatsiani, L., Low, T. Y. & Heck, A. J. R. Six alternative proteases for mass spectrometry-based proteomics beyond trypsin. *Nature Protocols* **11,** 993–1006 (2016).
8. Wang, Y. *et al.* Human PAD4 regulates histone arginine methylation levels via demethylimination. *Science* **306,** 279–283 (2004).
9. Mohammed, H. *et al.* Rapid immunoprecipitation mass spectrometry of endogenous proteins (RIME) for analysis of chromatin complexes. *Nature Protocols* **11,** 316–326 (2016).
10. Roux, K. J., Kim, D. I., Raida, M. & Burke, B. A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells. *J Cell Biol* **196,** 801–810 (2012).
11. Branon, T. C. *et al.* Efficient proximity labeling in living cells and organisms with TurboID. *Nature Biotechnology* **36,** 880–887 (2018).
12. Davies, S. P., Reddy, H., Caivano, M. & Cohen, P. Specificity and mechanism of action of some commonly used protein kinase inhibitors. *Biochem. J.* **351,** 95–105 (2000).
13. Mclauchlan, H., ELLIOTT, M. & Cohen, P. The specificities of protein kinase inhibitors: an update. *Biochem. J.* **371,** 199–204 (2003).
14. Bain, J. *et al.* The selectivity of protein kinase inhibitors: a further update. *Biochem. J.* **408,** 297–315 (2007).
15. Ying, Q.-L. *et al.* The ground state of embryonic stem cell self-renewal. *Nature* **453,** 519–523 (2008).
16. Mellacheruvu, D. *et al.* The CRAPome: a contaminant repository for affinity purification–mass spectrometry data. *Nature Methods* **10,** 730–736 (2013).
17. Darrah, E. *et al.* Erosive Rheumatoid Arthritis Is Associated with Antibodies That Activate PAD4 by Increasing Calcium Sensitivity. *Science translational medicine* **5,** 186ra65–186ra65 (2013).
18. Asaga, H., Nakashima, K., Senshu, T., Ishigami, A. & Yamada, M. Immunocytochemical localization of peptidylarginine deiminase in human eosinophils and neutrophils. *J Leukoc Biol* **70,** 46–51 (2001).
19. Nakashima, K., Hagiwara, T. & Yamada, M. Nuclear localization of peptidylarginine deiminase V and histone deimination in granulocytes. *J. Biol. Chem.* **277,** 49562–49568 (2002).
20. Cuthbert, G. L. *et al.* Histone Deimination Antagonizes Arginine Methylation. *Cell* **118,** 545–553 (2004).
21. Neeli, I. & Radic, M. Opposition between PKC isoforms regulates histone deimination and neutrophil extracellular chromatin release. *Front Immunol* **4,** 38 (2013).
22. Yap, K. L. *et al.* Calmodulin Target Database. *Journal of Structural and Functional Genomics* **1,** 8–14 (2000).
23. Kobertz, W. R., Mruk, K., Farley, B. M. & Ritacco, A. W. Predicting Calmodulin Binding Sites via Canonical Motif Clustering. *Biophysical Journal* **106,** 527a (2014).
24. Mruk, K., Farley, B. M., Ritacco, A. W. & Kobertz, W. R. Calmodulation meta-

analysis: predicting calmodulin binding via canonical motif clustering. *J. Gen. Physiol.* **144,** 105–114 (2014).
25. Abbasi, W. A., Asif, A., Andleeb, S. & Minhas, F. U. A. A. CaMELS: In silico Prediction of Calmodulin Binding Proteins and their Binding Sites. *Proteins: Structure, Function, and Bioinformatics* (2017). doi:10.1002/prot.25330
26. Rhoads, A. R. & Friedberg, F. Sequence motifs for calmodulin recognition. *FASEB J.* **11,** 331–340 (1997).
27. Slade, D. J. *et al.* Protein arginine deiminase 2 binds calcium in an ordered fashion: implications for inhibitor design. *ACS Chem. Biol.* **10,** 1043–1053 (2015).
28. Auger, I., Martin, M., Balandraud, N. & Roudier, J. Rheumatoid Arthritis-Specific Autoantibodies to Peptidyl Arginine Deiminase Type 4 Inhibit Citrullination of Fibrinogen. *Arthritis & Rheumatism* **62,** 126–131 (2010).
29. Chen, S. J. *et al.* Studies with Synthetic Peptide-Substrates Derived From the Neuronal Protein Neurogranin Reveal Structural Determinants of Potency and Selectivity for Protein-Kinase-C. *Biochemistry* **32,** 1032–1039 (1993).
30. Gallant, C., You, J. Y., Sasaki, Y., Grabarek, Z. & Morgan, K. G. MARCKS is a major PKC-dependent regulator of calmodulin targeting in smooth muscle. *J Cell Sci* **118,** 3595–3605 (2005).
31. Villalonga, P. *et al.* Calmodulin prevents activation of Ras by PKC in 3T3 fibroblasts. *J. Biol. Chem.* **277,** 37929–37935 (2002).
32. Agell, N. *et al.* The diverging roles of calmodulin and PKC in the regulation of p21 intracellular localization. *Cell Cycle* **5,** 3–6 (2006).
33. Faux, M. C. & Scott, J. D. Regulation of the AKAP79-protein kinase C interaction by Ca2+/calmodulin. *J. Biol. Chem.* **272,** 17038–17044 (1997).
34. Kruger, H. *et al.* Pkc Inhibitor From Bovine Brain Identified as Calmodulin. *Journal of Protein Chemistry* **7,** 257–258 (1988).
35. Nairn, A. C. & Aderem, A. *Calmodulin and Protein Kinase C Cross-Talk: The MARCKS Protein is an Actin Filament and Plasma Membrane Cross-Linking Protein Regulated by Protein Kinase C Phosphorylation and by Calmodulin*. *Ciba Foundation Symposium 164 - Interactions Among Cell Signalling Systems* **9,** 145–161 (John Wiley & Sons, Ltd, 2007).
36. Chan, K. F., Robb, N. D. & Chen, W. H. Myelin basic protein: interaction with calmodulin and gangliosides. *Journal of Neuroscience Research* **25,** 535–544 (1990).
37. Libich, D. S., Hill, C. M. D., Haines, J. D. & Harauz, G. Myelin basic protein has multiple calmodulin-binding sites. *Biochemical and Biophysical Research Communications* **308,** 313–319 (2003).
38. Kim, H., Jo, S., Song, H.-J., Park, Z.-Y. & Park, C.-S. Myelin basic protein as a binding partner and calmodulin adaptor for the BKCa channel. *Proteomics* **7,** 2591–2602 (2007).
39. Majava, V. *et al.* Interaction between the C-terminal region of human myelin basic protein and calmodulin: analysis of complex formation and solution structure. *BMC Struct. Biol.* **8,** 10–18 (2008).
40. Hammer, J. A., Jung, G. & Korn, E. D. Genetic evidence that Acanthamoeba myosin I is a true myosin. *PNAS* **83,** 4655–4659 (1986).
41. Jung, H. S. *et al.* Conservation of the regulated structure of folded myosin 2 in species separated by at least 600 million years of independent evolution. *PNAS* **105,** 6022–6026 (2008).
42. Odronitz, F. & Kollmar, M. Drawing the tree of eukaryotic life based on the analysis of 2,269 manually annotated myosins from 328 species. *Genome Biol.* **8,** R196–23 (2007).
43. Lee, K. H. *et al.* Interacting-heads motif has been conserved as a mechanism of myosin II inhibition since before the origin of animals. *PNAS* **115,** E1991–E2000 (2018).
44. Forrester, M. T., Foster, M. W., Benhar, M. & Stamler, J. S. Detection of protein S-nitrosylation with the biotin-switch technique. *Free Radical Biology and Medicine* **46,** 119–126 (2009).
45. Chung, H. S., Murray, C. I. & Van Eyk, J. E. A Proteomics Workflow for Dual Labeling

Biotin Switch Assay to Detect and Quantify Protein S-Nitroylation. *Methods Mol. Biol.* **1747,** 89–101 (2018).
46. Bauer, P. M., Buga, G. M., Fukuto, J. M., Pegg, A. E. & Ignarro, L. J. Nitric oxide inhibits ornithine decarboxylase via S-nitrosylation of cysteine 360 in the active site of the enzyme. *J. Biol. Chem.* **276,** 34458–34464 (2001).
47. Leiper, J., Murray-Rust, J., McDonald, N. & Vallance, P. S-nitrosylation of dimethylarginine dimethylaminohydrolase regulates enzyme activity: Further interactions between nitric oxide synthase and dimethylarginine dimethylaminohydrolase. *PNAS* **99,** 13527–13532 (2002).
48. Hao, G., Xie, L. & Gross, S. S. Argininosuccinate synthetase is reversibly inactivated by S-nitrosylation in vitro and in vivo. *J. Biol. Chem.* **279,** 36192–36200 (2004).
49. Hillary, R. A. & Pegg, A. E. Decarboxylases involved in polyamine biosynthesis and their inactivation by nitric oxide. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **1647,** 161–166 (2003).
50. Lai, T. S. *et al.* Calcium regulates S-nitrosylation, denitrosylation, and activity of tissue transglutaminase. *Biochemistry* **40,** 4904–4910 (2001).
51. Shirai, H., Blundell, T. L. & Mizuguchi, K. A novel superfamily of enzymes that catalyze the modification of guanidino groups. *Trends in Biochemical Sciences* **26,** 465–468 (2001).
52. Linsky, T. & Fast, W. Mechanistic similarity and diversity among the guanidine-modifying members of the pentein superfamily. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **1804,** 1943–1953 (2010).
53. Hardman, G. *et al.* Strong anion exchange mediated phosphoproteomics reveals extensive human non canonical phosphorylation. *The EMBO Journal* **13,** Unit 13 15–24 (2019).
54. Kizawa, K. *et al.* Specific Citrullination Causes Assembly of a Globular S100A3 Homotetramer: A Putative Ca2+ Modulatior Matures Human Hair Cuticle. *J. Biol. Chem.* **283,** 5004–5013 (2008).

Chapter 7: Thesis Summary

The five human Peptidyl arginine deiminases (PADIs) catalyse the PTM citrullination and have particular biomedical interest for their deregulation in diseases such as Rheumatoid Arthritis, Multiple Sclerosis and cancer. Chemical inhibition and genetic ablation have been shown to mitigate against these pathologies, making PADIs important therapeutic targets. Little however, is understood about how PADI activity is switched on in physiological contexts or how that becomes disrupted or overactive in disease. Work described in this thesis focused firstly on the evolutionary origin of the enzyme family in the human lineage showing the family of five mammalian PADI enzymes were acquired by a single ancient horizontal gene transfer event from cyanobacteria. I then looked at one of the human paralogues, PADI4, to address the critical open question as to how the enzyme may be regulated in cells. This identified calcium binding proteins which differentially interact with PADI4 protein before and after activation. One of these proteins, calmodulin, was shown to activate PADI4 in vitro. Lastly, the thesis focused on targeting PADI4 for inhibition, pulldown and activation with novel cyclic peptide reagents, performed in collaboration. Several cyclic peptides were shown to bind with high affinity to the human protein, and showed efficacy separately as enzyme inhibitors and as enzyme activators both in vitro and in cells.

Experiments in the first results chapter (Chapter 3) focused on the evolutionary origin of the PADI family. This work revealed there are a number of cyanobacterial, actinobacterial and fungal homologues in addition to the mammalian enzymes, but that, following phylogenetic analysis, the eukaryotic homologues separate into two distinct clades – one of actinobacterial and fungal homologues and one of cyanobacterial and animal homologues. Phylogenetic testing, analysis of synapomorphic protein features and domains, and Bayesian analysis of evolutionary rates and accumulated protein sequence changes show that the three domain animal homologues were introduced into the human lineage by ancient horizontal

gene transfer. The mammalian proteins derive from a late diverging clade of cyanobacteria that possess the closest degenerate three domain PADI ancestor and retain conservation of all six calcium binding sites in PADI2. Experiments showed the cyanobacterial protein is calcium dependent and catalytically active on mammalian substrates, targeting human histone 3, but possesses a different, even more stringent, calcium dependency than the mammalian enzyme. This points to exciting possibilities for future work to uncover how mechanisms for PADI regulation in the human lineage evolved.

Experiments in Chapter 4 and Chapter 6 looked at the activation of the human enzyme PADI4 in cells. In Chapter 4, cellular conditions were set up to activate human PADI4 in terminally differentiated HL-60 cells, in purified human neutrophils and in a stringently controlled model system related to the role of PADI4 in pluripotency. Cellular conditions were identified where PADI4 could be activated within a rapid time window under conditions that would be tractable for differential proteomic analysis. As part of this work, PADI4 was shown to be activated by canonical Wnt signalling using recombinant Wnt3a, different small molecule inhibitors of GSK3β and a chemical agonist of the Wnt pathway. Furthermore, inhibiting PADI4 reduced Wnt driven transcription in this system suggesting a new role for PADI4 in amplifying Wnt signalling. Future work will be required to verify this exciting signalling pathway is implicated in activating PADI4 in a physiologically relevant context such as in haematopoetic stem cells, or in cancer contexts.

In Chapter 6, proteomic experiments were undertaken to analyse PADI4 to identify differences between the inactive and activated enzyme. A number of novel PTMs were identified on PADI4 including a phosphorylation site close to the active site. Excitingly, a list of interacting proteins were also identified that differed in binding between the resting and active enzyme conditions – including a set of known calcium-binding proteins, which were revealed to preferentially bind PADI4 allosterically in the resting condition. For one of these proteins, calmodulin, the mode of binding was verified by reciprocal

pulldown and recombinant calmodulin was shown to activate human PADI4 in vitro by reducing the calcium dependency of the enzyme. This represents the first endogenous protein shown to activate PADI4 and provides a potential endogenous mechanism for the physiological activation of PADI4. Work in this chapter concluded with bioinformatic analysis to identify a putative calmodulin binding motif on human PADI4. A series of follow-up experiments to validate these effects in cells are left for future work.

Experiments in Chapter 5 complement the work in the thesis and describe a collaboration to target human PADI4 by screening a large library of cyclic peptides (>10^{13}) for tight binding affinity with human PADI4 with Dr Walport and Prof Suga. This was done using the Random Non-Standard Peptide Integrated Discovery (RaPID) system that was developed in the lab of Prof Suga with the hope to identify candidates that can inhibit, activate and pulldown human PADI4 from cells. Two screens were undertaken on the calcium-soaked active human enzyme conformation to enrich candidates firstly that bind at the active site and secondly that bind allosterically, by covalently blocking the active site prior to the screen. Biochemical and cellular experiments were undertaken and showed that separate cyclic peptides were able to inhibit as well as to activate PADI4. After sequence optimization, the PADI4 inhibitors and PADI4 activators show efficacy both in vitro and in cells. These represent novel reversible and potent PADI4 inhibitors, and the first to target the active enzyme conformation. The peptide activators represent the first such epigenetic activators. These reagents provide new tools for the citrullination field with use already demonstrated in the proteomic experiments in Chapter 6. In addition, they should enable specific questions to tease out the precise role for PADI4 activation in cellular contexts such as in innate immunity (in NET formation) and in the establishment of pluripotency.

This body of work presents progress towards the understanding of PADIs in human biology, with a focus on PADI4. An unexpected example of horizontal

gene transfer was identified as providing the evolutionary origin of the enzyme family. Work was undertaken to identify the precise events that modulate PADI4 activation implicating PADI4 activation as occurring downstream of the important Wnt signalling pathway and involving regulatory calcium-binding proteins, including calmodulin, in mediating enzyme activation. Finally, new chemical reagents for PADI4 were created to inhibit, activate and isolate the enzyme from cells, with efficacy in vitro and in cells. This PhD thesis presents progress towards understanding the critical question as to how PADI4 activation is achieved in cellular physiology and disease. Teasing out the complete mechanism for PADI4 activation will represent important follow-up work.

Appendix

A.1 Size exclusion chromatography methods

Lysis and methods for analyzing interacting complexes by native protein interaction size exclusion chromatography were devised in conversation with Dr Martin Wear, Edinburgh Protein Production Facility, University of Edinburgh. This was optimized for two 15 cm^3 dishes of mouse ES cells (approximately $20x10^6$ cells) per condition. 1mL of ice cold PBSGF (PBS, 5% glycerol, 0.5 mM DTT, 0.05% Tween-20, with protease inhibitors and PhosSTOP) were added (in total) directly to the two dishes and cells scraped with a silicon cell scraper. The two dishes were combined into a single Eppendorf, 2 µL benzonase and 2 mM $MgCl_2$ were added and the tube incubated at 4˚C for 40 minutes. Pellets were sheared by passing 10 times through a 25G needle and then centrifuged for 30 min at 25,000 x g at 4˚C. The lysate was analysed by immunoblot and revealed that sufficient hPADI4 protein was released under these lysis conditions. Proteins are then ready for loading onto a Superdex size exclusion column such that either 250 µL sample is loaded in a 500 µL loop (optimally), or 500 µL in a 1 mL loop. Columns are run at 0.5 mL/min with equilibration with at least 2CVs of buffer. Either 0.25 mL or 0.5 mL fractions can be collected across the elution profile. The first third of the elution will yield nothing, between 7 - 8 mL there is the void volume containing material that is too large to partition in the matrix before the range of dynamic separation from approximately 8 mL to 18 mL. Up to three wavelengths from 190 nm - 750 nm can be monitored, usually including 258 nm and 280 nm for protein and nucleotide and either a wavelength between 214 - 220 nm to analyse total peptide composition or 600 nm to assess the general turbidity of the sample. Fractions can be analysed by immunoblot, alongside a control protein, to see whether hPADI4 appears in a protein complex.

A.2 EMSA method

This electrophoretic mobility shift assay (EMSA) protocol was adapted from a protocol developed by Carlo De Angelis and who kindly provided initial

instruction. Three 20-24-mer oligo DNA probes are required, one forward strand (F), one reverse strand (R) and a third of either the forward or reverse which has been biotin labelled at the 3' end (either F* or R*). The third oligo can be biotin labelled using the Biotin 3' End labeling DNA kit (Thermo Scientifics, Prod no 89818) or synthesized with the label already attached (Sigma). Oligos were reconstituted to 100 µM in ddH2O. Probes were generated by incubating 10 µL of 1:500 diluted F* (biotin tagged), 1:500 diluted 10 µL R oligo, 10 µL H salt buffer (Roche/SureCut restriction enzyme buffer) in 70 µL H_2O at 95˚C for 5 min before allowing the reaction to equilibrate back to RT so that the oligos anneal. Competitor DNA was generated by incubating 20 µL of stock F, 20 µL of stock R oligo, 10 µL H salt buffer (Roche/SureCut restriction enzyme buffer) in 50 µL H_2O at 95˚C for 5 min before allowing the reaction to equilibrate back to RT so that the oligos anneal. This is diluted 1:5, such that 2 µL gives a 100 fold molar excess over probe DNA. Cellular nuclear extracts were prepared using NE-PER Nuclear and Cytoplasmic Extraction Reagents kit (Thermo Scientific 78833) with protease inhibitors added to yield a final approximate concentration of 2-4 mg/mL. 10X Binding buffer was prepared at 10mM Tris-HCl pH7.9, 50 mM KCl and 1 mM DTT and poly dI/dC (1mg/mL) was obtained from Sigma. Binding reactions without competitor DNA were performed by incubating 1.5 µL Binding Buffer, 1 µL poly dI/dC, 2 µL (4 µg) of nuclear extract, in 10.5 µL H2O at 4˚C for 5 min. Then 2 µL probe DNA (20 fmol) was added and the reaction incubated at RT for 20 min. Competitor binding reactions were performed by incubating 1.5 µL Binding Buffer, 1 µL poly dI/dC, 2 µL (4 µg) of nuclear extract, 2 µL annealed competitor DNA in 8.5 µL H2O at 4˚C for 5 min. Then 2 µL probe DNA (20 fmol) was added and the reaction incubated at RT for 20 min. A control of nuclear extract without oligo, and a control of oligo without nuclear extract were always included. Sometimes the nuclear extract can exhibit non-specific banding so it can be useful see the quantity of unbound oligo appearing at the bottom of the gel. After incubation, 4 µL of loading buffer is added, 30% sucrose in H_2O, Bromophenol Blue to colour.

Acrylamide gels (6%) were prepared with 4.5 mL 40% Acrylamide/Bis Acrylamide (19:1), 750 µL 20 x TBE, 90 µL 25% Ammonium persulfate, 35 µL TEMED up to 30 mL in H_2O (Sigma). The BioRad Mini-PROTEAN® 3 Cell (Catalog Numbers165-3301 165-3302) system was used. The 19 µL samples are loaded onto the gel. Rainbow marker (GE healthcare) is used to give an approximate indicator of band size. Gels were run at 100 V at 4˚C until the dye front reached the bottom of the gel without running off. The dye front is not discrete but runs broadly with trailing edges that correspond to the lanes. Gels were transferred to a positively charged nylon membrane (Roche) between 2 sheets of 3 M chromatography paper at 100 V for 25 min in a Biorad transfer apparatus using 0.5xTBE buffer with ice packs at 4˚C. Membranes were crossed linked before being processed using the Thermo Scientific Chemiluminescent Nucleic Acid detection Module (Product no. 89880). Membranes were blocked and then incubated with the Streptavidin-Horseradish Peroxidase Conjugate in heat-sealed polythene bags to reduce the quantity of reagents used. Membranes were washed in 0.5 x wash buffer (gives a cleaner result) and what follows is a recipe for homemade Equilibration buffer. The following equilibration buffer was used (10X buffer uses 1.2 g Tris, 0.58 g NaCl, 0.2 g $MgCl_2$, pH to 9.5 in 100 mL H_2O). Membranes were probed with the developer from the kit and exposed to Amersham Hyperfilm ECL. Blots took approximately 1-2 min to give sufficient signal.

A.3 PhosTag gel electrophoresis method

A protocol for Phos-tag gel electrophoresis adapted from *Ito et al*. and *Kinosita et al*. is presented[1,2]. Phosphorylated proteins visualised in the gel appear to migrate more slowly than corresponding dephosphorylated protein bands due to binding between phosphorylated amino acids with the Phos-tag reagent that is incorporated in the gel (obtained from Dr Hilary McLauchlan and Prof Dario Alessi at the MRC PPU). Following treatment, cells were washed with TBS (20 mM Tris-HCl, pH 7.5, and 150 mM NaCl) on ice and lysed in an ice-cold lysis buffer containing 50 mM Tris-HCl, pH 7.5, 1% (v/v)

Triton X-100 (NP-40), 1 mM EGTA, 1 mM sodium orthovanadate, 50 mM NaF, 0.1% (v/v) 2-mercaptoethanol, 10 mM 2-glycerophosphate, 5 mM sodium pyrophosphate, 0.1 µg/mL mycrocystin-LR (Enzo Life Sciences), 270 mM sucrose and protease inhibitors. Lysates were centrifuged at 17 000 x *g* for 20 min at 4°C and supernatants mixed with Loading buffer before being used for immunoblot analysis. Samples were supplemented with 10 mM $MnCl_2$ prior to loading gels.

Phos-tag gels were constructed with a 4% stacking gel (4% (w/v) acrylamide, 125 mM Tris-HCl, pH 6.8, 0.1% (w/v) SDS, 0.2% (v/v) *N,N,N′,N′*_tetramethylethylenediamine (TEMED), 0.08% (w/v) ammonium persulfate (APS)) layered above a 12% resolving gel (12% (w/v) acrylamide, 375mM Tris-HCl, pH 8.8**,** 0.1% (w/v) SDS, 75 µM Phos-tag acrylamide**,** 150 µM $MnCl_2$**,** 0.1% (v/v) *N,N,N′,N′*_tetramethylethylenediamine (TEMED)**,** 0.05% (w/v) ammonium persulfate (APS)). The gel mixture was degassed for 10 min before adding TEMED and APS (Sigma). After centrifugation at 17000 x *g* for 1 min, 30 µg of total protein were loaded per well and electrophoresed at 70 V for the stacking gel and at 150 V for the resolving gel using Phostag running buffer (25 mM Tris-HCl**,** 192 mM glycine**,** 0.1% (w/v) SDS). Gels were washed three times for 10 min in Phostag wash buffer (48 mM Tris-HCl**,** 39 mM glycine**,** 10 mM EDTA **,** 0.05% (w/v) SDS**,** 20% (v/v) methanol), followed by washing once for 10 min in Phostag transfer buffer (48 mM Tris-HCl**,** 39 mM glycine**,** 20% (v/v) methanol) supplemented with 0.05% SDS, but not with EDTA. Proteins were electrophoretically transferred onto nitrocellulose membranes at 100 V for 180 min on ice using Phostag transfer buffer (48 mM Tris-HCl**,** 39 mM glycine**,** 20% (v/v) methanol, without SDS or EDTA). Transferred membranes were blocked with 5% (w/v) BSA in TBS-T (20 mM Tris-HCl, pH 7.5, 150 mM NaCl**,** 0.1% (v/v) Tween 20) at room temperature for 30 min. Membranes were incubated with primary antibodies diluted in 5% BSA in TBS-T overnight at 4°C. Membranes were washed in TBS-T and then incubated with secondary antibodies (anti-rabbit-HRP abcam) diluted in 5% BSA in TBS-T at

room temperature for 1 h. Membranes were washed again in TBS-T and detection carried out by exposing films (Amersham Hyperfilm ECL (GE Healthcare) to the membranes using an ECL solution (SuperSignal West Dura Extended Duration (Thermo Fisher Scientific) or by imaging directly with the GE ImageQuant LAS 4000.

A.4 Cellular Thermal Shift Assay (CETSA) method

Per treatment condition, a T75 flask of cells was used –a treated condition was always compared to a vehicle only treatment. 10 mL of full serum media containing vehicle or inhibitor was added to the cells and incubated for 30 min or 1 hour. Following treatment, cells were washed in 6 mL PBS before incubate in 3 mL Accutase. After 3 min, 6 mL of full serum media was added and detached cells washed twice in PBS, before resuspension in PBS containing protease inhibitors. 100 µL of resuspended cells were added to 7 different tubes in a PCR strip. In a trial experiment PADI4 was found to be thermally unstable above approximately 56-59˚C. A PCR machine was used to heat the tube at a gradient between 46˚C and 70˚C. The PCR strip was heated in the middle lanes of a PCR block for 3 min with temperature thereby covering discontinuous steps from 50, 53, 56, 59, 62, 65, and 68 ˚C. The PCR strip was allowed to equilibrate back to RT for 5 min. Cell lysis was conducted by freezing in liquid nitrogen for 60 seconds and then thawing in a 25˚C water bath for 60 secs, with vortexing in between. This was repeated three times. After lysis, the contents of each PCR tube was transferred to a fresh Eppendorf and centrifuged at 17 000 x *g* for 20 min at 4˚C. 60 µL of supernatant was added to 20 µL 4X reducing loading buffer and identical quantities loaded for SDS PAGE gel running and Western blotting for human PADI4 with NPM1 or total H3 used as nuclear loading controls.

A.5 References for Appendices

1. Ito, G. et al. Phos-tag analysis of Rab10 phosphorylation by LRRK2: a powerful assay for assessing kinase function and inhibitors. *Biochem. J.* **473**, 2671–2685 (2016).
2. Kinoshita, E., Kinoshita Kikuta, E., Kubota, Y., Takekawa, M. & Koike, T. A Phos-tag SDS-PAGE method that effectively uses phosphoproteomic data for profiling the phosphorylation dynamics of MEK1. *Proteomics* (2016).